# Nonlinear Dynamic Field Embedding: On Hyperspectral Scene Visualization

Dalton Lunga and Okan Ersoy

## Abstract

Graph embedding techniques are useful to characterize spectral signature relations for hyperspectral images. However, such images consists of disjoint classes due to spatial details that are often ignored by existing graph computing tools. Robust parameter estimation is a challenge for kernel functions that compute such graphs. Finding a corresponding high quality coordinate system to map signature relations remains an open research question. We answer positively on these challenges by first proposing a kernel function of spatial and spectral information in computing neighborhood graphs. Secondly, the study exploits the force field interpretation from mechanics and devise a unifying nonlinear graph embedding framework. The generalized framework leads to novel unsupervised multidimensional artificial field embedding techniques that rely on the simple additive assumption of pair-dependent attraction and repulsion functions. The formulations capture long range and short range distance related effects often associated with living organisms and help to establish algorithmic properties that mimic mutual behavior for the purpose of dimensionality reduction. The main benefits from the proposed models includes the ability to preserve the local topology of data and produce quality visualizations *i.e.* maintaining disjoint meaningful neighborhoods. As part of evaluation, visualization, gradient field trajectories, and semisupervised classification experiments are conducted for image scenes acquired by multiple sensors at various spatial resolutions over different types of objects. The results demonstrate the superiority of the proposed embedding framework over various widely used methods.

## I. INTRODUCTION

**N**onlinear dimensionality reduction has emerged as a key preprocessing step for extracting

D. Lunga is with the Department of Electrical and Computer Engineering, Purdue University. He is also affiliated with the CSIR-Meraka Institute, Brummeria, Pretoria, South Africa e-mail: dlunga@purdue-edu.

O. Ersoy is with the Department of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, 47907-0501, USA e-mail: ersoy@purdue-edu.

and visualizing regular structures within complex data sets. Prior to its popularity, the easy to implement linear methods have spanned decades of research on this task dating back to principal component analysis(PCA) [1]–[3], the classical scaling or multidimensional scaling(MDS) technique [4], more recently the local Fisher discriminant analysis (LFDA) [5], and semi-supervised local Fisher discriminant analysis (SELF) [6]. With modern technology enabling the capability to gather and combine data from various sensing mechanisms, linear dimensionality reduction techniques have met their limitations in seeking meaningful structures from complex data [7]. For example, hyperspectral sensors have enabled the acquisition of greater details about objects on the earth surface which poses a challenge for linear dimensionality reduction techniques. Feature extraction in such data sets can be accomplished by employing nonlinear methods such as the maximum variance unfolding (MVU) [8]- a method that computes maximum variance embedding maps subject to preserving local distances, the locally linear embedding(LLE) [9] - a method that represents the relations of each neighborhood by linear coefficients that best reconstruct each data point from its neighbors, and the laplacian eigenmaps Laplacian(LE) [10], which draws on the correspondence between the graph Laplacian, the Laplace Beltrami operator on a manifold, and the connections to the heat equation, to devise a geometrically motivated algorithm for constructing a representation for data sampled from a low $m$-dimensional manifold embedded in a higher $d$-dimensional space.

Further widely popular methods include the generalized-MDS [11], [12] which extends the properties of classical scaling to nonlinear embedding, the self-organizing feature map ( SOFM) [13] which is an artificial unsupervised neural network for learning low-dimensional maps, graph embedding and extensions for dimensionality reduction [14]. Probabilistic approaches include the stochastic neighbor embedding(SNE) [15], a method that represents each object by a mixture of widely separated low dimensional factors capturing some of the local structure, and establishes global formations of clusters of similar maps. With improved modeling assumptions, variants of SNE have led to high quality embedding visualizations, *e.g.* the student $t$-distribution based stochastic neighbor embedding(tSNE) [16], the elastic embedding algorithm [17], and the spherical stochastic neighbor embedding( sSNE) [18]. In most cases, the approaches rely on high dimensional neighborhood graphs to capture the geometrical relations of observations and use graph weights as inputs for embedding purposes. The techniques are formulated under metric measures that include the geodesic distances, and for others, non-metric probability measures

including Kullback Leibler divergences.

Notwithstanding individual differences in efficiency and their specific data applications, nonlinear dimensionality reduction methods share some features including, better compression, better visualization, and reduction of classifier input features. Each embedding technique represents an attempt to search for a coordinate representation that resides on the data manifold. Recent comparative studies conducted on various nonlinear embedding techniques strongly indicated the inability of existing manifold learning methods to handle disjoint class structures that exists in hyperspectral data [19]. In addition, many dimensionality reduction algorithms suffer from what is known as the crowding problem in the machine learning community [16]. The crowding problem can be described as a tendency by which embedding techniques collapse maps towards the center of the embedding space resulting in increased class overlaps. Under such a phenomenon, many embedding algorithms fail to establish discriminative boundaries between structures of different classes. In this study several functional forms are proposed to address this challenge within a unifying framework for developing nonlinear dimensionality reduction algorithms with benefits that spans visualization, compression, and classification tasks.

While incorporating some benefits from existing methods, the study provides a unified multi-dimensional artificial force embedding( MAFE) framework for nonlinear dimensionality reduction with new perspectives and high quality embedding functional forms. MAFE models are based on seeking a minimum energy configuration state of a high dimensional neighborhood graph in a lower dimensional space. Underlying the framework is the premise that given a predefined graph structure, the optimal embedding coordinates should generate a corresponding low dimensional neighborhood graph that preserve the high dimensional pairwise relations. The embedding framework draws from the force field interpretation to suggest insightful design properties that lead to pair-dependent attraction and repulsion functions for composing novel potential embedding fields. The functions are superposed to generate an odd function that enforce the pairwise interaction fields in the embedding space. The generated interaction is such that at longer range the attraction force, which pulls similar maps towards each other, dominates, while the repulsion force dominates at short range distances. Under such an environment, all maps will experience a pushing and a pulling force from their relative neighbors due to the field generated by the potential functions. Of importance is to observe that the attraction function emphasizes the pulling together of similar maps while the repulsion function acts as a barrier

that generates a pushing force for maps to be dispersed. Much of the local relations that reveal meaningful structures in optimal embeddings should already have been captured by the similarity kernel function that constructs the neighborhood input graph. As a second contribution, the study combines MAFE models with a local bilateral similarity kernel for mapping spatial and spectral signatures onto the lower dimensional space. This is demonstrated by embedding real world remote sensing images that are characterized by high spectral resolution pixels with several hundred channels. However, having many channels and nonlinearities in each pixel poses several challenges to conventional land cover methods since visually different objects tend to exhibit overlapping or similar spectral signatures.

The paper is structured as follows. A review of the force field formulations is introduced in Section II. A general formulation of the multidimensional artificial field embedding(MAFE) framework is presented in Section III. In Section IV, we establish MAFE connections to existing popular nonlinear embedding techniques. In Section V, using MAFE framework, we propose two novel techniques: the multidimensional artificial field embedding - bounded repulsion(MAFE-BR), and the multidimensional artificial field embedding - unbounded repulsion(MAFE-UR). In Section VIII, we describe the experimental setup and results on various data sets. A discussion of experiments and future work is presented in Section IX. Finally, we present conclusions in Section X.

## II. FORCE FIELD MOTIVATION

Force field interpretations have a long history that relates to nature studies. In Biology, researchers have long studied nature and discovered that populations often appear in patterns of aggregation such as flocks of birds, schools of fish, and herds of mammals. Biological models that use forces between individuals that are analogous to physical forces, have since been developed. The study in [20], is an example of early work that proposed an idea of mutual interactions between individuals that were composed of attractions and repulsions with the goal of maintaining the group as a stable mass. This idea was developed further in [21] by discussing the possibility of modeling forces between individual fish upon classical gravitation and electromagnetism. In a later paper [22], the author considered inverse power laws to model the repulsions and attractions between individuals, with repulsion stronger at short inter-individual distances. The work in [22], was compared to actual data collected from schools of fish, to obtain model parameters. The

model was based on a simple constant attraction and a repulsion inversely proportional to the square of the inter-individual distance. The same laws governing animal behavior have been observed to be generalizable to guide artificial systems to carry out complex tasks by relying only on local interactions. In fact, in individual-based frameworks, the basis rule is that aggregation behavior is a result of an interplay between a short-ranged repulsion and a long-ranged attraction between the individuals. Such an intuition has been applied with success in control of multi-agent systems, stability analysis of social foraging swarms [23], and robotic motion planning [24].

Most Biology studies [20]–[22], [25], [26] and Control Engineering studies [23], [24], [27], seek to address questions that relates to maintaining a group as a stable mass, stability analysis, group cohesion and obstacle avoidance. Notwithstanding such growing recognition of the importance of force field formulation in those areas, their role in the allied area of signal and image processing remains to be fully appreciated. This study is most concerned on drawing on the collective dynamics of aggregation behavior to devise algorithms that establish manifold formations given a predefined structure, *i.e.* formation of multiple manifolds given a high dimensional neighborhood graph as a constraint. As such, we focus on the application of the force field intuition to the problem of dimensionality reduction and manifold learning based on a graph embedding optimization framework.

## III. Dynamic Graph Embedding Formulation

Let $\mathcal{G} = (\mathcal{E}, \mathcal{V})$ be a finite undirected graph with vertex set $\mathcal{V}$, edge set $\mathcal{E}$ and with no self loops. The elements of $\mathcal{E}$ are designated as ideal springs. Furthermore, let $\mathcal{S} = \{(w_{ij}, k_{ij})\}$ be the spring properties between each vertex $i$ and $j$ for all $\{(v_i, v_j)\} \in \mathcal{E}$, where $w_{ij}$ is the normalized or unnormalized length without compression or extension computed for each observed pair in $\boldsymbol{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_N\}$, where $\boldsymbol{y}_i \in \mathbb{R}^d$, and $k_{ij} = 1$ is the corresponding force constant. Let $\mathcal{G}_{\mathcal{S}}$ be a neighborhood spring graph, where $\mathcal{S}$ denotes the spring relation. An embedding of $\mathcal{G}_{\mathcal{S}}$ is an assignment of vertices into a $m$-dimensional Euclidean space $\mathbb{R}^m$. Let $\boldsymbol{Z} = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_N\}^T$ be the assigned embedding of $\mathcal{G}_{\mathcal{S}}$, where $\boldsymbol{z}_i \in \mathbb{R}^m$ is the position of vertex $i$'s map. When framed as a graph embedding task, where on each vertex the approach imagines a particle in motion and each edge weight is dictated by spring force laws, the corresponding problem of dimensionality reduction simply becomes that of establishing a minimum energy configuration that is governed

by the structure in $\boldsymbol{W} = [w_{ij}]$. The modeling assumptions are such that the configuration yields maps that preserve pairwise relations characterized by the neighborhood graph $\mathcal{G}_\mathcal{S}$. Finding such a mapping is at the heart of every dimensionality reduction model, and it is the subject that is discussed shortly.

A mechanics interpretation of the graph embedding framework is presented as follows. Imagine the existence of a particle on every $\boldsymbol{z}_i \in \boldsymbol{Z}$, that is moving with the velocity of $\boldsymbol{Z}$'s centroid. With the following change of notation to denote the embedding positions as a *state* of a graph, let $\mathcal{Z} = \left\{ \boldsymbol{z}_1^T, \boldsymbol{z}_2^T, \cdots, \boldsymbol{z}_N^T \right\}^T$ be a long vector in $\mathbb{R}^{Nm}$. Thus only consider the motion dynamics of individual maps, not the motion of the group. The approach assumes that all individual maps move simultaneously and each map $i$ is aware of the position of other vertices (maps), as well as the strength of corresponding force that defines each edge weight in the graph. The positions, $\boldsymbol{z}_i$'s, of individuals relative to the group centroid can change through the rearrangements informed by pair-dependent force field interactions. Assuming such motion is to change in a continuous time, then the velocity as determined by the effect of group members on each vertex $i$, at position $\boldsymbol{z}_i$ is described by

$$\dot{\boldsymbol{z}}_i = \sum_{j=1, j \neq i} F^{ij}(\boldsymbol{z}_i - \boldsymbol{z}_j), \quad i = 1, \cdots, N \tag{1}$$

where $F^{ij}(\boldsymbol{z}_i - \boldsymbol{z}_j) = (\boldsymbol{z}_i - \boldsymbol{z}_j) \left\{ F_r^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|) - F_a^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|) \right\}$ describes pairwise symmetric interactions between the $i$th and $j$th maps. Symmetry of the function follows from the fact that if map $i$ is attracted to map $j$, then $j$ is attracted to $i$. $F_r^{ij} : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ denotes the magnitude of the repulsion term, whereas $F_a^{ij} : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ represents the magnitude of the attraction term. The superposition of these two terms defines an interactive function that is the basis for the MAFE framework. Model insights and properties governing the choice of this function are presented in the following sections.

## A. Force Fields Function Properties

Artificial force field formulations assume that at large distances, the attraction function dominates, and that on short distances the repulsion function dominates, while in between there is a unique distance at which both terms will balance - defining a central path in a similar manner to the barrier method [28]. The choice of suitable force field embedding interaction functions, $F^{ij}(\boldsymbol{z}_i - \boldsymbol{z}_j)$, are guided by the following properties:

1) There is pair-equilibrium distance $\epsilon_{ij}$ at which $F_r^{ij}(\epsilon_{ij}) = F_a^{ij}(\epsilon_{ij})$, else $F_a^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|) > F_r^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|)$ for $\|\boldsymbol{z}_i - \boldsymbol{z}_j\| > \epsilon_{ij}$ or $F_a^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|) < F_r^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|)$ for $\|\boldsymbol{z}_i - \boldsymbol{z}_j\| < \epsilon_{ij}$.

2) $F^{ij}$ is an odd function, *i.e.* $F^{ij}(-(\boldsymbol{z}_i - \boldsymbol{z}_j)) = -F^{ij}(\boldsymbol{z}_i - \boldsymbol{z}_j)$, therefore symmetric with respect to the origin.

3) There exist pair dependent functions $U_{att}^{ij} \to \mathbb{R}^+ \to \mathbb{R}^+$ and $U_{rep}^{ij} \to \mathbb{R}^+ \to \mathbb{R}^+$ such that

$$\nabla_{\boldsymbol{z}_i} w_{ij} U_{att}^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|) = F_a^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|)(\boldsymbol{z}_i - \boldsymbol{z}_j)$$

$$\nabla_{\boldsymbol{z}_i} U_{rep}^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|) = F_r^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|)(\boldsymbol{z}_i - \boldsymbol{z}_j)$$

$U_{att}^{ij}$ and $U_{rep}^{ij}$ are viewed as artificial attraction and repulsion potential energy functions. The combined term $(\boldsymbol{z}_i - \boldsymbol{z}_j)F_r^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|)$ represents the actual repulsion effect, whereas the term $-(\boldsymbol{z}_i - \boldsymbol{z}_j)F_a^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|)$ represents the actual attraction effect. The vector $(\boldsymbol{z}_i - \boldsymbol{z}_j)$ establishes the alignment on which the attraction and repulsion interaction forces acts along in opposing directions. These functions describe the reactive approach by potential fields in which trajectories of particles motion are not planned explicitly. Instead the interactions of every map with its neighbors is a superposition of fields that enable its position to cope with the changing environment of other maps. The motion dynamics can be rewritten to reflect the resultant forces on each individual map as

$$\dot{\boldsymbol{z}}_i = -\sum_{j=1, j \neq i} \left\{ \nabla_{\boldsymbol{z}_i} w_{ij} U_{att}^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|) - \nabla_{\boldsymbol{z}_i} U_{rep}^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|) \right\}$$

The assumption made to envision each map as moving along the negative gradient has an implication, *i.e.* to achieve a minimum-energy configuration of the graph $\mathcal{G}_S$, a choice of the attraction and repulsion potential functions should be such that the minimum of $U_{att}^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|)$ occurs on or around $\|\boldsymbol{z}_i - \boldsymbol{z}_j\| = 0$, whereas the minimum of $-U_{rep}^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|)$ (or maximum of $U_{rep}^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|)$) occurs on or around $\|\boldsymbol{z}_i - \boldsymbol{z}_j\| \to \infty$, and that the minimum of the combination $U_{att}^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|) - U_{rep}^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|)$ occurs at $\|\boldsymbol{z}_i - \boldsymbol{z}_j\| = \epsilon_{ij}$, thus defining the stationery state of motion that exist between pairs $i$ and $j$.

Using the above framework, the reactive potentials that are effective on each individual map $i$ can be represented as

$$U_i(\mathcal{Z}) = \sum_{j=1, j \neq i} \left\{ w_{ij} U_{att}^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|) - U_{rep}^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|) \right\} \tag{2}$$

while the total superposed potential function on the neighborhood graph $\mathcal{G}_{\mathcal{S}}$ is defined by

$$U(\mathcal{Z}) = \sum_{i=1}^{N} U_i(\mathcal{Z}) \tag{3}$$

Letting $\Omega$ be the set of attraction and repulsion functions $F^{ij}(\cdot)$ satisfying the embedding field properties, new embedding models can simply be derived by solving the following general optimization problem:

$$\mathcal{Z}^{\star} = \underset{\mathcal{Z} \in \mathbb{R}^{Nm}}{\arg\min} U(\mathcal{Z}) \tag{4}$$

where $\mathcal{Z}^{\star}$ describes the minimum-energy configuration state of $\mathcal{G}_{\mathcal{S}}$ in the lower dimensional space. With some parameter adjustments on $F^{ij}(\cdot)$, the embedding maps will converge to a minimum-energy configuration that yields the optimal maps. Such an embedding framework can be adapted as a general platform to develop new nonlinear dimensionality reduction algorithms but first, we establish its links to existing popular embedding techniques in the following section.

## IV. Connections To Existing Methods

Many nonlinear techniques have been proposed for the tasks of visualization and dimensionality reduction. Even though with differences, they are applied in various fields as preprocessing building blocks for compression, visualization, and classification tasks. A basic question that we ask is whether some of the existing methods can be derived as special cases of the MAFE framework? A positive illustration to this question is presented together with reformulations of popular techniques including the stochastic neighbor embedding(SNE) [15], student-t stochastic neighbor embedding [16], Laplacian eigenmaps (LE) [10], Cauchy graph embedding [29], and the spherical stochastic neighbor embedding(sSNE) [18]. Such interpretations show that MAFE is a unifying framework that not only inherits some algorithmic benefits of various techniques, but also provides extended functional properties combined with strong intuitive insights for creating new algorithms.

### A. Stochastic Neighbor Embedding

Stochastic neighbor embedding [15] is a method for preserving probabilities on lower dimensional manifolds that are nonlinear. SNE assumes that edge weights are antisymmetric probabilities $w_{ij}$ (i.e. $w_{ij} \neq w_{ji}$) of pairs of vertices being neighbors in the higher dimensional

space. However, our presentation focuses on the symmetric version where $w_{ij} = w_{ji}$ for all pairs of vertices. The high dimensional edge weights are defined using the Gaussian functions of the form,

$$w_{ij} = \frac{\exp\{-\frac{\|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2}{2\sigma_i}\}}{\sum_{r=1,r\neq i} \exp\{-\frac{\|\boldsymbol{y}_r - \boldsymbol{y}_i\|^2}{2\sigma_i}\}} \tag{5}$$

where $\sigma_i$ is computed using a binary search method ensuring that the entropy of the distribution $W_i$ is approximately $\log(k)$, with $k$ defining the effective number of neighbors. In the lower dimensional space, a symmetric Gaussian probability $\hat{w}_{ij}$ is assumed between each pair of embedding maps, *i.e.* the embedding graph weights are computed as

$$\hat{w}_{ij} = \frac{\exp\{-\|\boldsymbol{z}_i - \boldsymbol{z}_j\|^2\}}{\sum_{r=1,r\neq i} \exp\{-\|\boldsymbol{z}_r - \boldsymbol{z}_i\|^2\}} \tag{6}$$

Each $\boldsymbol{z}_i \in \mathbb{R}^m$ is the corresponding lower dimensional map of the observation $\boldsymbol{y}_i \in \mathbb{R}^d$ . SNE proceeds to compute for the maps by minimizing a sum of Kullback Leibler(KL) objective functions

$$\sum_i KL(W_i \| \hat{W}_i) = \sum_i \sum_{j=1,j\neq i} w_{ij} \log(\frac{w_{ij}}{\hat{w}_{ij}}) \tag{7}$$

The goal of (7) is to minimize the distortion between each of the $N$ high dimensional neighborhood distributions $W_i$ and their corresponding lower dimensional neighborhood distributions $\hat{W}_i$. The results obtained from this approach have so far demonstrated its superiority when compared to methods that include locally linear embedding(LLE) [9], MDS [4], and Isomap [30]. However, the optimization algorithm is very unstable which leads to a lot of experimentally defined parameters in order to attain meaningful results. A further expansion on (7) while ignoring terms that are not a function of the lower dimensional maps (terms that do not depend on $\hat{w}_{ij}$), much in parallel to the work of [17], reveals the log-sum term as a source of difficulty when computing the gradient and its a term that increases the nonlinearity of the model.

Computing the negative gradient of (7) yields the corresponding MAFE motion dynamics or force field equations of the form

$$\dot{\boldsymbol{z}}_i^{SNE} = -\nabla_{\boldsymbol{z}_i} U^{SNE} = -4 \sum_{r=1,r\neq i} F^{SNE}(\boldsymbol{z}_i - \boldsymbol{z}_j) \tag{8}$$

where the expression under summation is defined as

$$F^{SNE}(\boldsymbol{z}_i - \boldsymbol{z}_j) = (\boldsymbol{z}_i - \boldsymbol{z}_j)\left\{w_{ij} - \frac{\exp\{-\|\boldsymbol{z}_i - \boldsymbol{z}_j\|^2\}}{\sum_{r=1,r\neq i} \exp\{-\|\boldsymbol{z}_r - \boldsymbol{z}_i\|^2\}}\right\} \tag{9}$$

Under the MAFE formulation, equation (8) describes the motion state of vertex maps seeking the graph's minimum energy configuration in the lower dimensional space. A simple observation identifies the attractive gradient force field to be $-w_{ij}(z_i - z_j)$, and a repulsion gradient force field to be $\frac{(z_i - z_j)\exp\{-\|z_i - z_j\|^2\}}{\sum_{r=1,r\neq i}\exp\{-\|z_r - z_i\|^2\}}$. The interpretation of (8) is such that, at longer distances with parameters set carefully, the embedding maps start to form clusters due to the strong attraction force field, while the repulsion force field is very negligible. As $\|z_i - z_j\| \to 0$ for each $(i,j)$ pair, the repulsion magnitude dominates the interaction force vector. This causes the maps to push away from each other. The convergence of the algorithm is established when the forces balance, much in the same way as described for the general MAFE framework.

## B. t-Stochastic Neighbor Embedding

$t$-Stochastic Neighbor Embedding [16] is similar to SNE except that the lower dimensional maps are assumed to be better modeled by a *Student t-distribution* of degree one. This simple modification leads to a complete improvement of results over SNE. The improvement is due to the pair-dependent inverse distance relation introduced by the Student t-distribution. Using this distribution, the joint embedding probabilities $\hat{w}_{ij}$ are defined by

$$\hat{w}_{ij} = \frac{(1 + \|z_i - z_j\|^2)^{-1}}{\sum_{r=1,r\neq i}(1 + \|z_r - z_i\|^2)^{-1}} \tag{10}$$

In tSNE, the cost function is also based on the sum of Kullback Leibler(KL) divergences (same as in (7)). The corresponding negative gradient of the expanded tSNE cost function gives the following MAFE motion dynamics or force field equations,

$$
\begin{aligned}
\dot{z}_i^{tSNE} &= -\nabla_{z_i} U^{tSNE} \\
&= -4 \sum_{r=1,r\neq i} \left\{ w_{ij}\frac{(z_i - z_j)}{1 + \|z_i - z_j\|^2} - \frac{(z_i - z_j)(1 + \|z_i - z_j\|^2)^{-2}}{\sum_{r=1,r\neq i}(1 + \|z_r - z_i\|^2)^{-1}} \right\} \\
&= -4 \sum_{(i,j)\in\mathcal{E}} F^{tSNE}(z_i - z_j) \tag{11}
\end{aligned}
$$

where the term under summation in (11) is defined as

$$F^{tSNE}(z_i - z_j) = (z_i - z_j) \left\{ \frac{w_{ij}}{1 + \|z_i - z_j\|^2} - \frac{(1 + \|z_i - z_j\|^2)^{-2}}{\sum_{r=1,r\neq i}(1 + \|z_r - z_i\|^2)^{-1}} \right\} \tag{12}$$

In a similar manner, equation (11) describes the motion state of vertex maps seeking the graph's minimum energy configuration in the lower dimensional space. However, the attractive force field

is described by $-(\boldsymbol{z}_i - \boldsymbol{z}_j)\frac{w_{ij}}{1+\|\boldsymbol{z}_i-\boldsymbol{z}_j\|^2}$, a term whose magnitude approaches an inverse square law for large pairwise distances $\|\boldsymbol{z}_i - \boldsymbol{z}_j\|$. This property makes the coordinate representation of joint probabilities invariant to changes in scale for maps that are far-apart. The repulsion force field is described by $(\boldsymbol{z}_i - \boldsymbol{z}_j)\frac{(1+\|\boldsymbol{z}_i-\boldsymbol{z}_j\|^2)^{-2}}{\sum_{r=1,r\neq i}(1+\|\boldsymbol{z}_r-\boldsymbol{z}_i\|^2)^{-1}}$. The magnitude of the inverse square law approximation by both terms dictates that formation of clusters be established at long range distances, while the repulsion force field is magnified at short range distances, causing all maps to disperse from each other. The convergence of the algorithm is established when the two force fields balance.

## C. sPherical Stochastic Neighbor Embedding

A spherical stochastic neighbor embedding model [18] is also a variant of SNE, whose embedding representations are constrained to reside on a spherical manifold. As such, it assumes the joint probability distribution $\hat{W}_i$ to be defined on a spherical surface, $\mathbb{S}^m = \{\boldsymbol{z} \in \mathbb{R}^{m+1} : \|\boldsymbol{z}\|_2 = 1\}$, with its corresponding probable values $\hat{w}_{ij}$ obtained by means of an Exit distribution [31]. sSNE, similarly to SNE, exhibits a desirable property that enables the unfolding of many-to-one mappings from which the same class objects are in several disparate locations that are spatially driven. The probable values $\hat{w}_{ij}$ are computed from the Exit expression

$$\hat{w}_{ij} = \frac{\frac{(1-\varrho)}{\|\boldsymbol{z}_j-\varrho\boldsymbol{z}_i\|^m}}{\sum_{r=1,r\neq i}\frac{(1-\varrho)}{\|\boldsymbol{z}_r-\varrho\boldsymbol{z}_i\|^m}} \tag{13}$$

where $\varrho$ is a concentration parameter that controls the assignment of neighborhood weights for a given $\boldsymbol{z}_i^{th}$ central map. sSNE's total cost function given by

$$\sum_i^N KL(W_i\|\hat{W}_i) + \lambda(\boldsymbol{z}_i^T\boldsymbol{z}_i - 1) = \sum_i \sum_{j=1,j\neq i} w_{ij}\log(\frac{w_{ij}}{\hat{w}_{ij}}) + \lambda(\boldsymbol{z}_i^T\boldsymbol{z}_i - 1) \tag{14}$$

where $\lambda$ is the Lagrangian parameter enabling the implicit incorporation of the unit spherical constraint in the objective function. The negative gradient of the expanded sSNE function yields the corresponding MAFE motion dynamic equations of the form

$$\begin{aligned}
\dot{\boldsymbol{z}}_i^{sSNE} &= -2m\varrho \sum_{j=1,j\neq i} (\boldsymbol{z}_j - \varrho\boldsymbol{z}_i)\left\{\frac{w_{ij}}{\|\boldsymbol{z}_j-\varrho\boldsymbol{z}_i\|^2} - \frac{(\|\boldsymbol{z}_j-\varrho\boldsymbol{z}_i\|)^{-m-2}}{\sum_{r=1,r\neq i}\|\boldsymbol{z}_r-\varrho\boldsymbol{z}_i\|^{-m}}\right\} - 2\lambda\boldsymbol{z}_i \\
&= 2m\varrho \sum_{j=1,j\neq i} F^{sSNE}(\boldsymbol{z}_i - \boldsymbol{z}_j) - 2\lambda\boldsymbol{z}_i \tag{15}
\end{aligned}$$

where the pairwise gradient force field on each map is defined by $F^{sSNE}(\cdot) = F^E(\cdot) - 2\lambda z_i$, with

$$F^E(z_i - z_j) = -(z_j - \varrho z_i) \left\{ \frac{w_{ij}}{\|z_j - \varrho z_i\|^2} - \frac{(\|z_j - \varrho z_i\|^{-m-2})}{\sum_{r=1, r \neq i} \|z_r - \varrho z_i\|^{-m}} \right\} \tag{16}$$

Contrary to both SNE and tSNE, equation (15) describes a MAFE dynamic model with maps constrained to be on a spherical manifold and in turn defining a unit sphere neighborhood graph configuration. The attractive force field is described by $-(z_j - \varrho z_i)\frac{w_{ij}}{\|z_j - \varrho z_i\|^2}$, a term whose magnitude is based on the inverse unbounded square law for large pairwise distances $\|z_i - z_j\|$. This property also makes the spherical coordinate representation of joint probabilities invariant to changes in scale for maps that are far-apart. The repulsion force field is described by $(z_j - \varrho z_i)\frac{(\|z_j - \varrho z_i\|^{-m-2})}{\sum_{r=1, r \neq i} \|z_r - \varrho z_i\|^{-m}}$, also an unbounded inverse distance term. The unbounded magnitude of the inverse square law by both terms enables sSNE to exhibit much needed capability to introduce a splitting nature on clusters that overlap. However, due to the unbounded nature on both terms, as $\|z_j - \varrho z_i\| \to 0$, the minimum energy configuration of the graph becomes unstable. This introduces algorithmic inefficiencies when seeking the optimal embedding coordinates for large data sets [18].

*D. Laplacian and Cauchy Embedding*

All graph embedding techniques presented in this study make the assumption that if observation $i$ and $j$ are similar, *i.e.* $w_{ij}$ is large, then the proximity of their corresponding maps $z_i$ and $z_j$ in the embedding space should also be closer. The $m$-dimensional embedding coordinates can also be obtained using the Laplacian eigenmaps (LE) approach [10]. Considering the 1-dimensional example, the idea is to minimize

$$\underset{z \ s.t. \|z\|^2 = 1, z^T \mathbf{1} = 0}{\operatorname{argmin}} U^{LE}(z) = \sum_{j=1, j \neq i} w_{ij}(z_i - z_j)^2 \tag{17}$$

where $\mathbf{1} = [1, \cdots, 1]^T$ and $\|z\|^2 = 1$ is a magnitude constraint imposed to avoid obtaining a solution where all $z_i$'s are zero, whereas the constraint $z^T \mathbf{1} = 0$, reduces the uncertainty on the non-uniqueness of the embedding solution. Under such constraints the problem is equivalent to

$$\underset{z}{\operatorname{argmin}} U^{LE}(z) = z^T L z \tag{18}$$

where $L = D - W$ is the graph Laplacian, $D = \operatorname{diag}(d_1, \cdots, d_N), d_i = \sum_j w_{ij}$ and $W = [w_{ij}]$ is the edge weight matrix. The solution is often computed by solving equation (18) as an

eigenvector problem. A simple observation on (17) connects the Laplacian eigenmaps formulation to the MAFE framework. A force field interpretation of the Laplacian embedding identifies, $\sum_{j=1,j\neq i} w_{ij}(\boldsymbol{z}_i - \boldsymbol{z}_j)^2$, as the artificial attractive potential whose corresponding negative artificial force field provides the motion equation for maps as defined by

$$\dot{\boldsymbol{z}}_i^{LE} = -\frac{dU^{LE}(\boldsymbol{z})}{dz_i} = -4 \sum_{j=1,j\neq i} F^{LE}(z_i - z_j) \tag{19}$$

where $F^{LE}(\boldsymbol{z}_i - \boldsymbol{z}_j) = (\boldsymbol{D} - \boldsymbol{W})_{ij}(z_i - z_j)$. The dynamic Laplacian model of (19) does incorporate the magnitude and non-uniqueness constraints from its objective function. However, these constraints do not generate a repulsion field that can mitigate the crowding of embeddings. The Cauchy graph embedding [29] approach has a MAFE interpretation that follows the same steps as applied in reformulation of Laplacian eigenmaps.

## V. NEW MAFE IMAGE EMBEDDING MODELS

Apart from the key insights obtained from the force field properties to design suitable pair-depended attraction and repulsion potential energies for a MAFE embedding model, another essential component involves choosing a similarity function suitable for incorporating all local information that is relevant for computing a neighborhood graph that captures the local structure of high dimensional observations. In most popular techniques, the Gaussian kernel described in equation (5) is used to compute such relations. However, for certain application areas with spatial information, *e.g.* hyperspectral imagery, the Gaussian weights may not be suitable. As such, a parametric local kernel for computing the spectral neighborhood is discussed in the next section. The parametric kernel can be seen as an extension of bilateral filter function [32]–[34] with a robust high dimensional covariance estimation component that enhances pixel differences and generates a sparse neighborhood graph. A discussion on the design of practical and better kernel functions tend to be application-driven and it remains an open research problem. Its complete investigation is therefore not fully addressed in this study.

### A. Spectral Neighborhood Graph Similarities

In various image processing applications, spatial preprocessing methods are often applied to remove noise and smooth images. These methods also enhance spatial texture information resulting in features that improve the performance of classification techniques. For example

in [35], nonlinear diffusion partial differential equations (PDEs) and wavelet shrinkage were used for spatial preprocessing of hyperspectral images, and the results obtained demonstrated a significant improvement on classification performance. Unlike the use of PDEs, we adapt a *local bilateral filtering* approach to devise a pairwise pixel similarity function or kernel over the observed neighborhood graph. The local kernel function is defined by

$$w(\boldsymbol{s}_i, \boldsymbol{s}_j, \boldsymbol{y}_i, \boldsymbol{y}_j) = \exp\left\{\frac{-\|\boldsymbol{s}_i - \boldsymbol{s}_j\|^2}{\sigma_s^2}\right\} \cdot \exp\left\{-(\boldsymbol{y}_i - \boldsymbol{y}_j)^T \boldsymbol{\Sigma}_y^{-1}(\boldsymbol{y}_i - \boldsymbol{y}_j)\right\} \tag{20}$$

where $\boldsymbol{s}_i$ denotes the spatial coordinates of pixel $i$, $\boldsymbol{y}_i$ denotes the photometric $d$-dimensional spectral vector, with $d$ corresponding to the number of spectral channels. The expression $\|\boldsymbol{s}_i - \boldsymbol{s}_j\|^2$ weighs image pixel values as a function of the spatial distance from the center pixel and $h_s$ is the variance parameter. The kernel also employs a nonlinear term,

$$\tilde{w}_p(i, j) = \exp\left\{-\frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{y}_j)^T \boldsymbol{\Sigma}_y^{-1}(\boldsymbol{y}_i - \boldsymbol{y}_j)\right\} \tag{21}$$

which simply weighs pixel values as a function of the photometric differences between the center pixel and its neighbor pixels. For example, given $N$ hyperspectral pixels, organized into a zero-mean data matrix $\boldsymbol{Y} = [\boldsymbol{y}_1 \boldsymbol{y}_2 \cdots \boldsymbol{y}_N] \in \mathbb{R}^{d \times N}$, the sample covariance is computed as $\boldsymbol{S} = \frac{1}{N}\boldsymbol{Y}\boldsymbol{Y}^T = \langle \boldsymbol{y}\boldsymbol{y}^T \rangle$, with the angle brackets denoting the average over $N$ pixels. Thus, $\boldsymbol{S}$ is a $d \times d$ matrix whose diagonal components indicate the magnitude of noise variation in each of the $d$ spectral channels, and the off-diagonal elements denote the extent to which noise co-vary with each pair of spectral bands. It is easy to show [18] that the photometric weights can be summarized by

$$\tilde{w}_p(i, j) = \exp\left\{\frac{-N}{2}tr(\boldsymbol{\Sigma}^{-1}\boldsymbol{S})\right\} \tag{22}$$

We make an observation that one can represent the unnormalized kernel $\mathcal{K}$ as a product of unnormalized gaussian functions, one for each pixel $\boldsymbol{y}_i$, yielding

$$\begin{aligned}
\mathcal{K} &= \exp\left\{\frac{-1}{2}\sum_{j=1}^{N}(\boldsymbol{y}_j - \boldsymbol{y}_i)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_j - \boldsymbol{y}_i)\right\} \\
&= \exp\left\{\frac{-tr(\boldsymbol{S}\boldsymbol{\Sigma}^{-1})}{2} + \sum_{j=1}^{N}\boldsymbol{y}_j\boldsymbol{\Sigma}^{-1}\boldsymbol{y}_i - \frac{N}{2}\boldsymbol{y}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y}_i\right\}
\end{aligned}$$

where $\boldsymbol{\Sigma}^{-1}$ is the inverse covariance matrix, and $tr(\boldsymbol{B})$ denotes the trace of matrix $\boldsymbol{B}$. We could assume a zero-mean unnormalized Gaussian noise model over the pixels, i.e. we can simply

subtract the center $\boldsymbol{y}_i$ from the data, to obtain a simplified expression as

$$\begin{aligned} \mathcal{K} &= \exp\left\{ \frac{-1}{2} \sum_{j=1}^{N} \boldsymbol{y}_j^T \boldsymbol{\Sigma}^{-1} \boldsymbol{y}_j \right\} \\ &= \exp\left\{ \frac{-1}{2} tr(\boldsymbol{Y}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{Y}) \right\} \end{aligned} \tag{23}$$

Note that $tr(\boldsymbol{Y}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{Y}) = tr(\boldsymbol{\Sigma}^{-1} \boldsymbol{Y}^T \boldsymbol{Y}) = N tr(\boldsymbol{\Sigma}^{-1} \boldsymbol{S})$. This shows that $\boldsymbol{S}$ is a sufficient statistic for characterizing the unnormalized likelihood (herein the photometric similarity) of data $\boldsymbol{Y}$, and we can further write

$$\mathcal{K} = \exp\left\{ \frac{-N}{2} tr(\boldsymbol{\Sigma}^{-1} \boldsymbol{S}) \right\} \tag{24}$$

In practice the weights are computed by first decomposing the true covariance matrix into a product $\boldsymbol{\Sigma} = \boldsymbol{E} \boldsymbol{\Lambda} \boldsymbol{E}^T$, where $\boldsymbol{E}$ is the orthogonal eigenvector matrix and $\boldsymbol{\Lambda}$ is the corresponding diagonal matrix of eigenvalues, that easily compute the covariance matrix whose inverse is required. We adapt the efficient sparse matrix transform (SMT) approach in estimating the covariance matrix $\boldsymbol{\Sigma}$ [36]. The SMT approach solves the optimization problem, $\hat{\boldsymbol{E}} = \operatorname{argmin}_{\boldsymbol{E} \in \boldsymbol{\Omega}} \left\{ |\operatorname{diag}(\boldsymbol{E}^T \boldsymbol{S} \boldsymbol{E})| \right\}$, and set $\hat{\boldsymbol{\Lambda}} = \operatorname{diag}(\hat{\boldsymbol{E}}^T \boldsymbol{S} \hat{\boldsymbol{E}})$, where $\boldsymbol{\Omega}$ is the set of allowed orthogonal transforms that can be computed using a series of *Givens rotations* [36]. A simple manipulation can show that $\boldsymbol{\Sigma}^{-1} = \hat{\boldsymbol{E}} \hat{\boldsymbol{\Lambda}}^{-1} \hat{\boldsymbol{E}}^T$ so that we have

$$\tilde{w}_p(i,j) = \exp\left\{ -\frac{1}{2} (\hat{\boldsymbol{E}}^T \boldsymbol{y}_i - \hat{\boldsymbol{E}}^T \boldsymbol{y}_j)^T \hat{\boldsymbol{\Lambda}}^{-1} (\hat{\boldsymbol{E}}^T \boldsymbol{y}_i - \hat{\boldsymbol{E}}^T \boldsymbol{y}_j) \right\}$$

with the final graphs weights computed from

$$w(\boldsymbol{s}_i, \boldsymbol{s}_j, \boldsymbol{y}_i, \boldsymbol{y}_j) = \exp\left\{ \frac{-\|\boldsymbol{s}_i - \boldsymbol{s}_j\|^2}{\sigma_s^2} \right\} \cdot \tilde{w}_p(i,j) \tag{25}$$

The SMT approach to computing the covariance matrix $\boldsymbol{\Sigma}$ is efficient and robust in handling the singularities of $\boldsymbol{\Sigma}$. Other approaches to computing $\boldsymbol{\Sigma}$ have been used in the literature including the PCA adaptation approach [32], where the singularity of $\boldsymbol{\Sigma}$ is not carefully addressed.

### B. Attractive Potential Function

The main fundamental idea underpinning the design of multidimensional artificial field embedding models is to treat the pair-equilibrium distances $\epsilon_{ij}$ for vertices in $\mathcal{G}_S$ as attractive wells. That is, consider the minimum-energy configuration between maps $i$ and $j$ as a sink for the potential energy function. For example, the attractive potential energy functions that we consider

are bounded from below to allow for the existence of constant attraction force effects at all distances, that is $U_{att}^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|) \geq \alpha$, where $\alpha$ is a positive constant $\forall \ \|\boldsymbol{z}_i - \boldsymbol{z}_j\|$. In this study, we explore potential functions of the form

$$U_{att}^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|) = \xi_a \|\boldsymbol{z}_i - \boldsymbol{z}_j\|^p \tag{26}$$

where for values $0 < p \leq 1$, the pairwise function is conic in shape, and the resulting attractive force field has a constant cluster formation amplitude determined from the graph's edge weights $w_{ij}$. $\xi_a$ is an attraction force magnitude related adaptive parameter. Figure 2 shows the attractive potential energy generated from equation (26) for $p = 2, \xi_a = 1$, and $w_{ij} = 0.5$. The shape corresponds to quadratic potential, *i.e.* we have a global optimal that acts to pull all force fields effects in its direction, thereby demonstrating the sink nature of the minimum point.

### C. Repulsive Potential Functions

The nature of a repulsion function is chosen such that as the distance between pair-points increases, its properties are deemed to have negligible influence on maps (*i.e.* maps are in long range zone where $F_r^{ij} < F_a^{ij}$). However, when the distance is small, the function generates a barrier or a repulsive force between maps (*i.e.* maps are in the short range zone where $F_r^{ij} > F_a^{ij}$). The procedure to selecting a repulsive potential function $U_{rep}^{ij}$ starts by thinking of an indicator function of the form

$$I_+(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|) = \begin{cases} 0 & \|\boldsymbol{z}_i - \boldsymbol{z}_j\| > \epsilon_{ij} \\ \infty & \|\boldsymbol{z}_i - \boldsymbol{z}_j\| \leq \epsilon_{ij}. \end{cases} \tag{27}$$

which is a non-increasing function of distance. Equation (27) best captures the behavioral form for a repulsive function that has negligible effects at large distances and has a large dominance at short range distances. With all its appealing intuitive properties, the indicator function is not a differentiable function. As such, we require its approximation by a differentiable function whose gradient can create a repulsion force $F_r^{ij}$ with a magnitude that is inversely proportional to the distance between pairs of maps *i.e.* $\|F_r^{ij}(\boldsymbol{z}_i - \boldsymbol{z}_j)\|_2 = \frac{1}{\mathbf{dist}(\boldsymbol{z}_i, \boldsymbol{z}_j)}$. Such approximations can be chosen from *e.g.* Gaussian, Exponential, Cauchy, Hyperbolic Tangent and Inverse distance power functions. The behavior of such functions in approximating $I_+(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|)$ is shown in Figure 1.
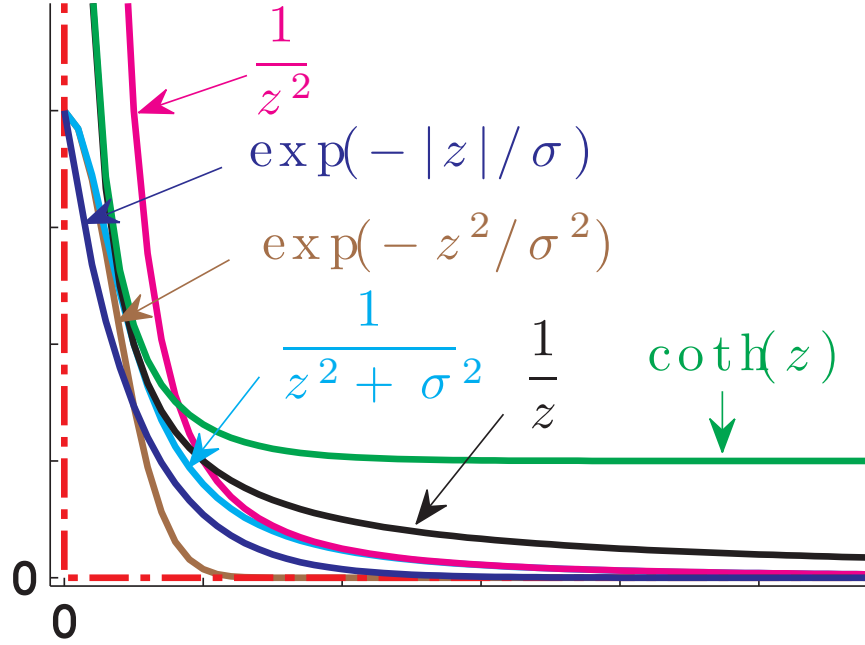
Fig. 1. Dashed lines show the function $I_+(\boldsymbol{z})$, and the solid curves show different forms of continuous decaying functions suitable for approximating $I_+(\boldsymbol{z})$.

*1) Exponential Bounded Repulsion:* The Exponential or unnormalized Gaussian curves in Figure 1 generates a continuous bounded approximation of (27). As the distance between maps grows the magnitude of the curves approaches zero while a maximum magnitude is assigned for maps that are very close in distance. A general unnormalized Gaussian bounded repulsion function can be devised in the form,

$$U_{rep}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|) = \xi_r \sigma \exp\{-\frac{\|\boldsymbol{z}_i - \boldsymbol{z}_j\|^q}{\sigma}\} \tag{28}$$

Where $\sigma_r$ is the bounded repulsion variance parameter. For $q = 2$, the function has spherical symmetry as shown in Figure 2. For values $0 < q \le 1$ the repulsion potential field has the shape of a harmonic function often used in modeling obstacles in robotic path planning [24], while for $1 < q < 2$ it has the form of a tower centered at the origin to generate equidistributed repulsive force fields in the direction of the gradient as shown in Figure 2.

*2) Inverse Power Unbounded Repulsion:* The inverse distance function as shown in Figure 1 has a continuous best approximation of the indicator function in equation (27). Its corresponding

general repulsion potential function is given by,

$$U_{rep}^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|) = \frac{\xi_r}{\|\boldsymbol{z}_i - \boldsymbol{z}_j\|^q} \qquad (29)$$

where $q$ is a positive number. It is clear that as $\|\boldsymbol{z}_i - \boldsymbol{z}_j\| \to 0^+$ the repulsion becomes unbounded, i.e. for $q = 2$, $\lim_{\|\boldsymbol{z}_i - \boldsymbol{z}_j\| \to 0^+} U_{rep}^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|)\|\boldsymbol{z}_i - \boldsymbol{z}_j\| = \infty$. The inverse square law structure of this function promotes an invariant representation of neighborhood similarities even with a change in scale for embedding points that are far apart. $\xi_r$ is an repulsion magnitude related parameter. Figure 3 shows the unbounded repulsion potential field generated from equation (29). In sharp contrast to existing embedding methods, this function affords properties that exhibit collision avoidance of maps as their coordinates change in search for the pair-equilibrium distances $\epsilon_{ij}$. The strong repulsive force field generated by this function renders it more favorable for rejoining of clusters that may have split during the search for a minimum energy configuration state of $\mathcal{G}_{\mathcal{S}}$.

### D. Multidimensional Artificial Field Embedding with Bounded Repulsion

A combination of the attractive and bounded unnormalized Gaussian repulsion functions yields a multidimensional artificial field embedding (MAFE-BR) model described by

$$U(\mathcal{Z}) = \sum_{i=1}^{N} \sum_{j=1, j \neq i} \left\{ W_{ij} U_{att}^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|) - U_{rep}^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|) \right\} \qquad (30)$$

$$= \sum_{i=1} \sum_{j=1, j \neq i} \left\{ \xi_a w_{ij} \|\boldsymbol{z}_i - \boldsymbol{z}_j\|^p - \xi_r \sigma \exp\{-\frac{\|\boldsymbol{z}_i - \boldsymbol{z}_j\|^q}{\sigma}\} \right\} \qquad (31)$$

For $q = 2$, (31) yields a special case elastic embedding model [17]. Computing the gradient provides information related to the direction and magnitude of motion for each individual map in the embedding space. This motion is described by

$$\dot{\boldsymbol{z}}_i = -\sum_{j=1, j \neq i} (\boldsymbol{z}_i - \boldsymbol{z}_j) \left\{ \xi_a w_{ij} p \|\boldsymbol{z}_i - \boldsymbol{z}_j\|^{p-2} - \xi_r q \|\boldsymbol{z}_i - \boldsymbol{z}_j\|^{q-2} \exp\{-\frac{\|\boldsymbol{z}_i - \boldsymbol{z}_j\|^q}{\sigma}\} \right\}$$

An illustration of the gradient field for a point with strong attraction force field is shown in Figure 2.
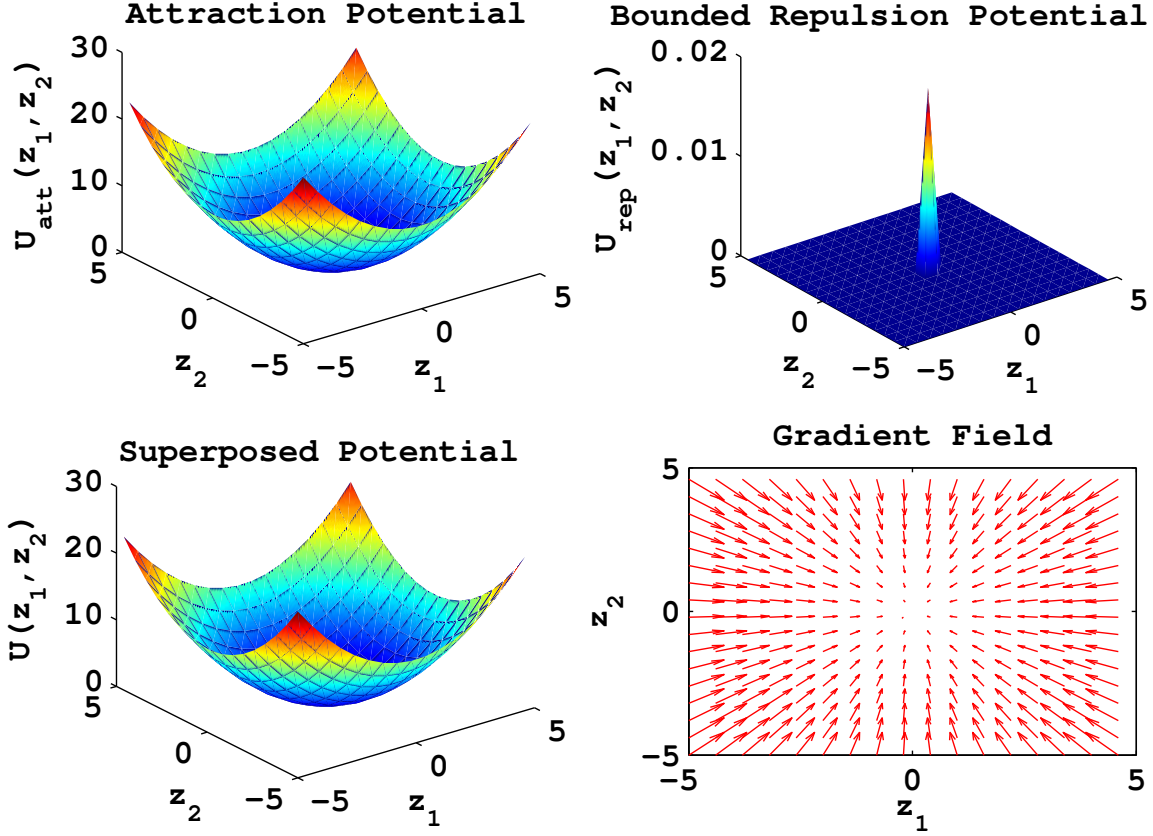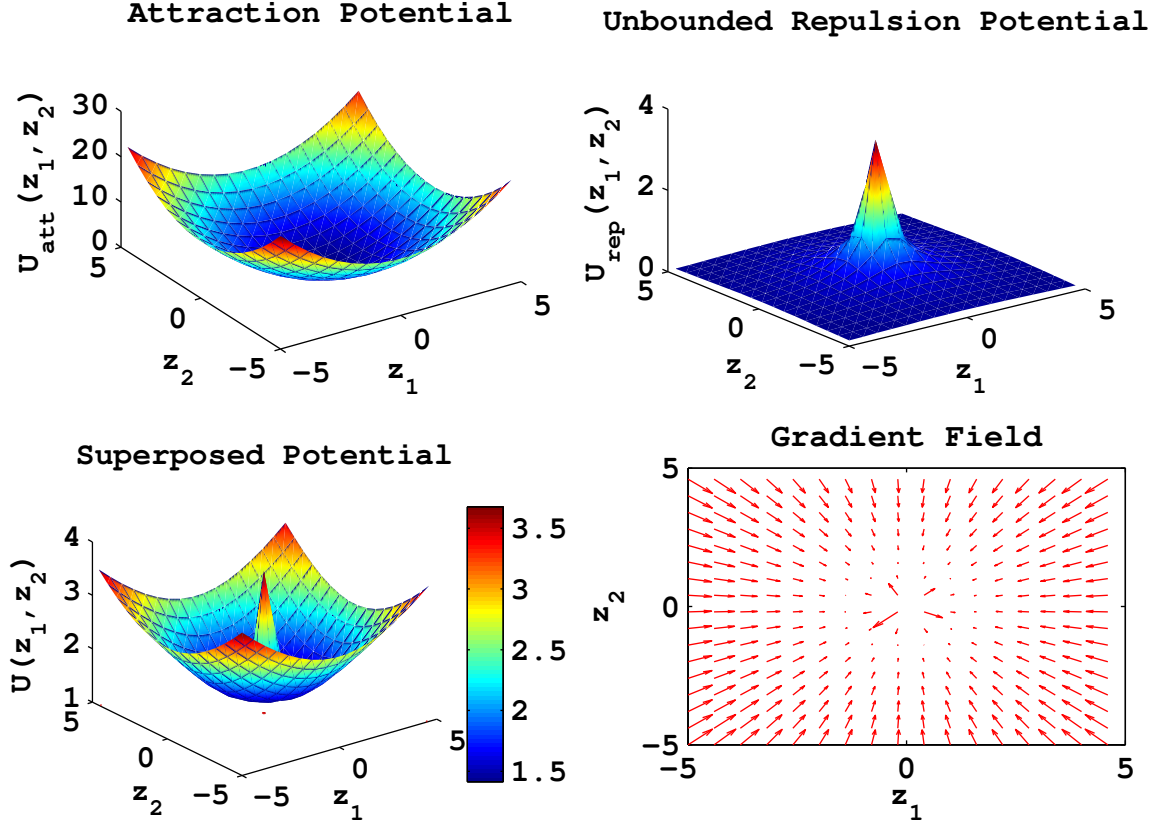
Fig. 2. Illustrations of a superposed-potential function for $p = q = 2$. Arrows indicate the negative gradient force field. The map located at $(z_1, z_2) = (0, 0)$ has a very strong attraction force over a large range of distance. The repulsion force is significantly small and mostly effective over a very short range.

### E. Multidimensional Artificial Field Embedding with Unbounded Repulsion

The second model that is proposed entails a multidimensional artificial field embedding model with an unbounded repulsion inverse distance function of (29). By considering all pairwise map interactions for vertices of $\mathcal{G}_\mathcal{S}$, the resultant total potential function of a MAFE-UR model is given by

$$
\begin{aligned}
U(\mathcal{Z}) &= \sum_{i=1} \sum_{j=1, j \neq i} \left\{ w_{ij} U_{att}^{ij}(\|z_i - z_j\|) - U_{rep}^{ij}(\|z_i - z_j\|) \right\} \\
&= \sum_{i=1} \sum_{j=1, j \neq i} \left\{ \xi_a w_{ij} \|z_i - z_j\|^p - \frac{\xi_r}{\|z_i - z_j\|^q} \right\}
\end{aligned}
\tag{32}
$$

Fig. 3. Illustrations of a superposed-unbounded repulsion potential function for $p = 1, q = 2$. Arrows indicate the negative gradient force field. The map located at $(z_1, z_2) = (0, 0)$ has a very strong repulsion force over on short range distances, while the attraction force is dominant over long range distances.

Under the supposed artificial force field model, the motion of each map is given by $\dot{z}_i = -\nabla_{z_i} U(z)$, i.e.

$$
\begin{aligned}
\dot{z}_i &= -\sum_{j=1, j \neq i} \left\{ \nabla_{z_i} w_{ij} U_{att}^{ij}(\|z_i - z_j\|) - \nabla_{z_i} U_{rep}^{ij}(\|z_i - z_j\|) \right\} \\
&= -2 \sum_{j=1, j \neq i} (z_i - z_j) \left\{ \xi_a w_{ij} p \|z_i - z_j\|^{p-2} + \xi_r q \|z_i - z_j\|^{-q-2} \right\}
\end{aligned}
\tag{33}
$$

Fig. 3, shows the total potential field generated from equation (32) and its corresponding gradient field obtained from (33) for $p = 1, q = 2$.

An important observation on the gradient fields in Figure 2 and Figure 3 demonstrates that without an attraction term in each model, cluster formation would not occur since all pair-wise maps would disperse from each other; whereas by eliminating the repulsion term(by setting

TABLE I

A SUMMARY OF FUNCTIONS COMPATIBLE TO DESIGN MULTIDIMENSIONAL ARTIFICIAL FORCE FIELD EMBEDDING
TECHNIQUES.

| Technique | Long Range Attraction | Short Range Repulsion | |
|---|---|---|---|
| SNE | $\sum_{i,j=1}^{N} w_{ij}\|\boldsymbol{z}_i - \boldsymbol{z}_j\|^2$ | $\sum_{i=1}^{N} \log \sum_{j=1}^{N} e^{\{-\|\boldsymbol{z}_i - \boldsymbol{z}_j\|^2\}}$ | $\sum_j w_{ij} = 1$ |
| tSNE | $\sum_{i,j=1}^{N} w_{ij} \log(1 + \|\boldsymbol{z}_i - \boldsymbol{z}_j\|^2)$ | $\sum_{i=1}^{N} \log(\sum_{j=1}^{N} \frac{1}{1+\|\boldsymbol{z}_i - \boldsymbol{z}_j\|^2})$ | $\sum_j w_{ij} = 1$ |
| SSNE | $\sum_{i,j=1}^{N} w_{ij} \log(\|\boldsymbol{z}_i - \rho\boldsymbol{z}_j\|^m)$ | $\sum_{i=1}^{N} \log(\sum_{j=1}^{N} \frac{1}{\|\boldsymbol{z}_i - \rho\boldsymbol{z}_j\|^m})$ | $\boldsymbol{z}_i \in \mathbb{S}^m$ |
| LE | $\sum_{i,j=1}^{N} w_{ij}\|\boldsymbol{z}_i - \boldsymbol{z}_j\|^2$ | $\boldsymbol{Z}\boldsymbol{Z}^T = I, \ \boldsymbol{Z}\boldsymbol{1} = \boldsymbol{0}$ | |
| CE | $\sum_{i,j=1}^{N} \frac{w_{ij}}{\sigma^2 + \|\boldsymbol{z}_i - \boldsymbol{z}_j\|^2\}}$ | $\boldsymbol{Z}\boldsymbol{Z}^T = I, \ \boldsymbol{Z}\boldsymbol{1} = \boldsymbol{0}$ | |
| GE | $\sum_{i,j=1}^{N} w_{ij} \exp\{-\|\boldsymbol{z}_i - \boldsymbol{z}_j\|^2\}$ | $\boldsymbol{Z}\boldsymbol{Z}^T = I, \ \boldsymbol{Z}\boldsymbol{1} = \boldsymbol{0}$ | |
| MAFEBR | $\sum_{i,j=1}^{N} w_{ij}\xi_a\|\boldsymbol{z}_i - \boldsymbol{z}_j\|^p$ | $\sum_{i,j=1}^{N} \xi_r\sigma \exp\{-\frac{\|\boldsymbol{z}_i - \boldsymbol{z}_j\|^q}{\sigma}\}$ | $p, q \geq 1$ |
| MAFEUR | $\sum_{i,j=1}^{N} w_{ij}\xi_a\|\boldsymbol{z}_i - \boldsymbol{z}_j\|^p$ | $\sum_{i,j=1} \frac{\xi_r}{\|\boldsymbol{z}_i - \boldsymbol{z}_j\|^q}$ | $p, q \geq 1$ |
| MAFEE | $\sum_{i,j=1}^{N} w_{ij}\sigma_a \exp\{-\frac{\|\boldsymbol{z}_i - \boldsymbol{z}_j\|^p}{\sigma_a}\}$ | $\sum_{i,j=1}^{N} \xi_r\sigma_r \exp\{-\frac{\|\boldsymbol{z}_i - \boldsymbol{z}_j\|^q}{\sigma_r}\}$ | $p, q \geq 1$ |
| MAFEH | $\sum_{i,j=1}^{N} \frac{w_{ij}\xi_a}{\|\boldsymbol{z}_i - \boldsymbol{z}_j\|^p}$ | $\sum_{i,j=1} \frac{\xi_r}{\|\boldsymbol{z}_i - \boldsymbol{z}_j\|^q}$ | $p, q \geq 1$ |

$\xi_r = 0$), all maps would collapse to a single point leading to the crowding problem that has been a weakness in most existing embedding models. Table I summarizes typical force field functions that can be used to design embedding models that exhibit aggregation behavior.

## VI. STOCHASTIC GRADIENT DESCENT OPTIMIZATION

The objective functions for MAFE based models are by far less nonlinear as compared to SNE and tSNE, as such their optimization is relatively simple, requiring no random jitter terms to establish stability. The optimization adapts a variation of the stochastic gradient descent [37] with a common adaptive learning rate

$$\alpha^{(t+1)} = \alpha^{(t)} + \gamma_1 \langle \nabla U(\mathcal{Z}^{(t-1)}), \nabla U(\mathcal{Z}^{(t)}) \rangle + \gamma_2 \langle \nabla U(\mathcal{Z}^{(t-2)}), \nabla U(\mathcal{Z}^{(t-1)}) \rangle \qquad (34)$$

where $\alpha^{(t)}$ is the learning rate at iteration $t$, $\gamma_1$ and $\gamma_2$ are the meta-learning rates. The main characteristics of this fast learning rate adaptation scheme is that it exploits gradient-related information from the current as well as the two previous embedding coordinates in the sequence. This provides an enhancement on the stabilization in the values of the learning rate and helps the gradient descent algorithm to exhibit even fast convergence that leads to better minimum energy-configuration. A description of the proposed algorithm is given in Algorithm 1. The algorithm's termination condition is when $\|\nabla U(\mathcal{Z})\| \leq \epsilon$. The choice of $\gamma_1$ and $\gamma_2$ is not critical for finding

the minimum-energy configuration but only affect the rate at which the algorithm does so. Figure 6 shows gradient field trajectories that were generated by this optimization scheme for a real world data set.

---

**Algorithm 1:** Generalized MAFE Algorithm

---

**Input**: High dimensional observations $\boldsymbol{Y}$ ;

Initialize optimization parameters: $\alpha^{(1)}$, $\gamma_1$, $\gamma_2$;

**Output**: Embedding coordinates matrix $\boldsymbol{Z} = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_N\}$;

Compute high dimensional neighborhood graph weights $w_{ij}$ from equation (25) or use a Gaussian kernel;

Randomly sample initial maps from a normal distribution *i.e.* $\boldsymbol{Z}^{(0)} \sim N(0, 50I)$;

Set $\mathcal{Z}^{(1)} = [\boldsymbol{z}_1^{(0)T}, \boldsymbol{z}_2^{(0)T}, \cdots, \boldsymbol{z}_N^{(0)T}]^T \in \mathbb{R}^{Nm}$ ;

**while** $\|\nabla U(\mathcal{Z}^{(t)})\| > \epsilon$ **do**

> Set $t = t + 1$;
>
> Compute new embedding coordinates using;
>
> $\mathcal{Z}^{(t+1)} = \mathcal{Z}^{(t)} - \alpha^{(t)}\nabla U(\mathcal{Z}^t)$;
>
> Calculate the new learning rate from
>
> $\alpha^{(t+1)} = \alpha^{(t)} + \gamma_1 \langle \nabla U(\mathcal{Z}^{(t-1)}), \nabla U(\mathcal{Z}^{(t)}) \rangle + \gamma_2 \langle \nabla U(\mathcal{Z}^{(t-2)}), \nabla U(\mathcal{Z}^{(t-1)}) \rangle$;

**end**

---

The iterative optimization in (4), or in particular the models described by equations (32), and (33) converges when all pair-equilibrium distances $\epsilon_{ij}$ are established. A strong theoretical argument can be made to assert that the motion of maps is guaranteed to stop and that no oscillatory behavior exist at the minimum-energy configuration state. This is done by letting the invariant set of the equilibrium positions to be

$$\Xi_{equi} = \left\{ \mathcal{Z} : \dot{\mathcal{Z}} = 0 \right\}.$$

The proof needs to show that as $t \to \infty$ the state $\mathcal{Z}(t)$ converges to $\Xi_{equi}$, *i.e.* the minimum-energy configuration of a graph converges to a stationery state. This extends Theorem 2 of [23] to problems of data visualization and dimensionality reduction.

**Theorem 1.** *Consider a graph embedding described by* $\dot{\boldsymbol{z}}_i = \sum_{j=1, j\neq i} F^{ij}(\boldsymbol{z}_i - \boldsymbol{z}_j)$, $i = 1, \cdots, N$, *with pairwise force field functions*

$F^{ij}(\boldsymbol{z}_i - \boldsymbol{z}_j) = (\boldsymbol{z}_i - \boldsymbol{z}_j)\{F_r^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|) - F_a^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|)\}$. *As $t \to \infty$, it can be shown that* $\mathcal{Z}(t) \to \Xi_{equi}$.

*Proof:* Consider the general energy function $U(\mathcal{Z}) = \sum_{i=1}^{N} U_i(\mathcal{Z})$, where $U_i(\mathcal{Z})$ is defined in (2). Taking the derivative of $U(\mathcal{Z})$ with respect to each $\boldsymbol{z}_i$ yields,

$$\nabla_{\boldsymbol{z}_i} U(\mathcal{Z}) = 2 \sum_{j=1, j \neq i} \left\{ \nabla_{\boldsymbol{z}_i} w_{ij} U_{att}^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|) - \nabla_{\boldsymbol{z}_i} U_{rep}^{ij}(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|) \right\} = -\dot{\boldsymbol{z}}_i$$

where the negative gradient is observed as direction of motion in the second equality. Taking the time derivative of $U(\mathcal{Z})$ along the motion of a graph configuration yields,

$$\dot{U}(\mathcal{Z}) = \nabla_{\boldsymbol{z}_i} U(\mathcal{Z})^T \dot{\mathcal{Z}} = 2 \sum_{i=1}^{N} \nabla_{\boldsymbol{z}_i} U(\mathcal{Z})^T \dot{\boldsymbol{z}}_i = 2 \sum_{i=1}^{N} \{-\dot{\boldsymbol{z}}_i\}^T \dot{\boldsymbol{z}}_i = -2 \sum_{i=1}^{N} \|\dot{\boldsymbol{z}}_i\|^2 \leq 0, \quad \forall t.$$

This result shows that the motion will continue in the direction of decreasing $U(\mathcal{Z})$ to a state when all $\dot{\boldsymbol{z}}_i = 0$. By invoking the Lasalle Invariance Principle [38], it can be concluded that as $t \to \infty$ the graph configuration state $\mathcal{Z}(t)$ converges to the largest subset of the set defined as

$$\Xi = \left\{ \mathcal{Z} : \dot{U}(\mathcal{Z}) = 0 \right\} = \{ \mathcal{Z} : \dot{\boldsymbol{z}}_i = 0 \} = \Xi_{equi}.$$

Since each $\dot{\boldsymbol{z}}_i \in \Xi_{equi}$ is an equilibrium point, $\Xi_{equi}$ is an invariant set and this concludes the proof. ∎

This general result holds for any function $F^{ij}$ chosen based on the embedding force field properties discussed in Section III. As such, it guarantees the convergence of our algorithms. However, in practise, the termination condition is set to some $\epsilon$ finite optimization of $U(\mathcal{Z})$. The result also extends to SNE, tSNE, sSNE and other related MAFE reformulations.

In addition to obtaining optimal embedding coordinates, the optimization of MAFE-BR and MAFE-UR with $p = q = 2$, also learns the embedding space neighborhood graph weights. For MAFE-BR, the optimal embedding neighborhood graph is described by

$$\tilde{w}_{ij}^{br} = \xi_a w_{ij} - \xi_r \exp\{-\frac{\|\boldsymbol{z}_i - \boldsymbol{z}_j\|^2}{\sigma}\}$$

The iterative optimization of MAFE-UR generates embedding graph weights that are given by

$$\tilde{w}_{ij}^{ur} = \xi_a w_{ij} - \frac{\xi_r}{\|\boldsymbol{z}_i - \boldsymbol{z}_j\|^4}$$

In contrast to the high dimensional neighborhood graph weights , $w_{ij}$, that are positive, the lower dimensional graph weights $\tilde{w}_{ij}^{br}$, and $\tilde{w}_{ij}^{ub}$, corresponding to the optimal embedding coordinates

(a)



(b)
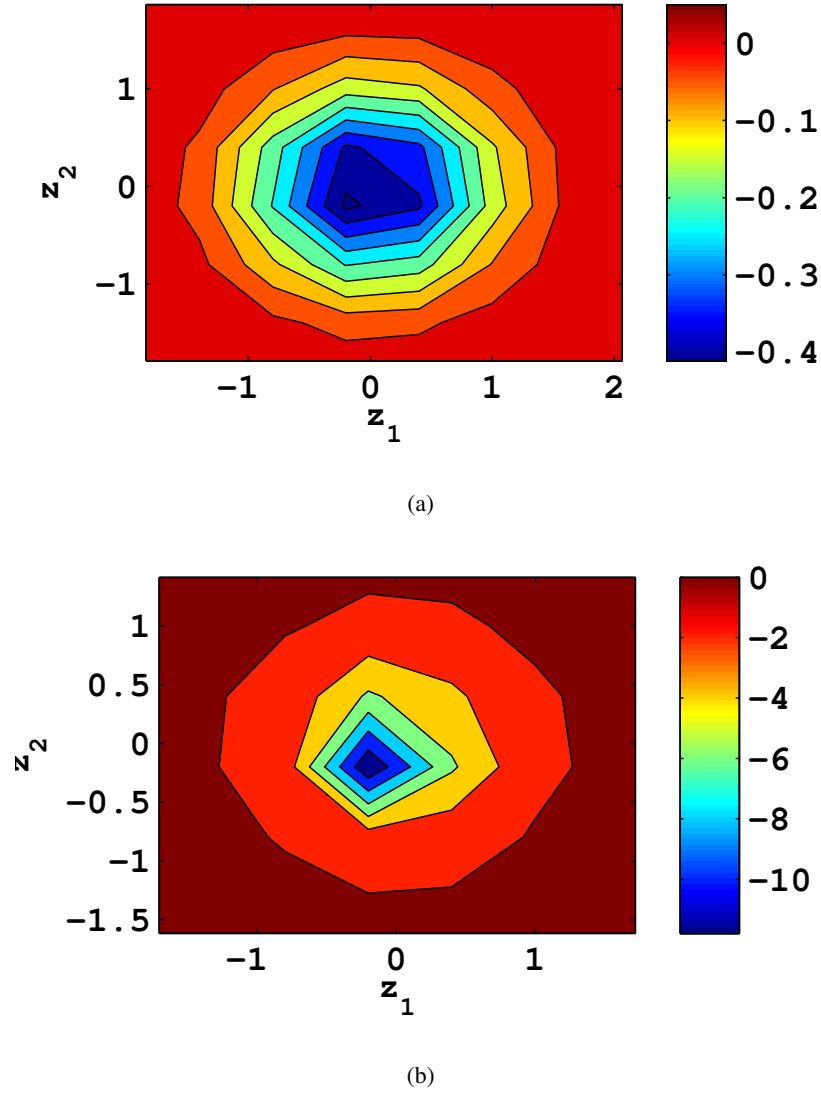
Fig. 4. Minimum energy graph configuration weights that are learned after 100 iterations for a KSC *Graminoid Marsh* pixel map with coordinates $\boldsymbol{z} = [-0.02, -0.003]$. MAFE-BR weights, $\tilde{w}_{ij}^{br}$, are shown in (a) and MAFE-UR weights, $\tilde{w}_{ij}^{ur}$, are shown in (b). For both models, the search for optimal embedding coordinates also leads to learning of negative graph weights.

can be negative as illustrated in Figure 4. This property establishes another point of significant contrast between the proposed MAFE based techniques in comparison to traditional embedding techniques that are known to enforce learning of positive lower dimensional weights *e.g.* LLE, SNE, tSNE, and Isomap.

## VII. Hyperspectral Image Scenes

*1) Botswana Hyperion:* Hyperion data with nine identified classes of complex natural vegetation were acquired over the Okavango Delta, Botswana, in May 2001, [39]. The general class groupings include seasonal swamps, occasional swamps, and woodlands. Signatures of several classes are spectrally overlapped, typically resulting in poor classification accuracies. After removing water absorption, noisy, and overlapping spectral bands, 145 bands were used for classification experiments. Classification error rates on signatures are carried out after data has been embedded to a lower dimensional space.

*2) Kennedy Space Center (KSC):* Airborne hyperspectral data were acquired by the National Aeronautics and Space Administration(NASA) Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) sensor at 18-m spatial resolution over Kennedy Space Center during March 1996. Noisy and water absorption bands were removed, leaving 176 features for thirteen wetland and upland classes of interest. Cabbage Palm Hammock (Class 3) and Broad Leaf/Oak Hammock (Class 6) are upland trees; Willow Swamp (Class 2), Hardwood Swamp (Class 7), Graminoid Marsh (Class 8) and Spartina Marsh (Class 9) are trees and grasses in wetlands. Their spectral signatures are mixed and often exhibit only subtle differences. Visualization and classification results for all thirteen classes are reported using the lower dimensional space coordinates.

*3) Indian Pine:* This scene was gathered by an AVIRIS sensor over the Indian Pines test site in North-western Indiana in June 12, 1992 at 20-m spatial resolution. The 220-band image scene is over an area that is 6 miles west of West Lafayette. The image has 16 classes consisting of agricultural fields that are alfalfa, corn-notill, cornmin, corn, grass/pasture, grass/trees, grass/pasture-mowed, haywindrowed, oats, soybeans-notill, soybean-min, soybeanclean, wheat, woods, dldg-grass-tree-drives, and stone-steel towers. The image size is $145 \times 145$ pixels. The pixel resolution is 16 bits, corresponding to 65536 gray levels. 2550 pixels were selected to generate the sample training and testing data. Again, we report on Euclidean and non-Euclidean embedding results as well as evaluation of the computed coordinates based on classification error rates for all 16 classes.

*4) Salinas-A:* This is a small sub-scene of Salinas image, denoted Salinas-A, that is commonly used in hyperspectral data analysis. It comprises $86 \times 83$ pixels located within the same scene at [samples, lines] = [591-676, 158-240] and includes six classes. The original scene was collected by the 224-band AVIRIS sensor over Salinas Valley, California, and is characterized by high

TABLE II

GROUND TRUTH CLASSES FOR THE SALINAS-A AND INDIAN PINE SCENES WITH THEIR RESPECTIVE SAMPLES NUMBERS

| Salinas-A | | Indian Pine | |
|---|---|---|---|
| c1 | Brocoli-green-weeds-1 (391) | c1 | Alfalfa (54) |
| c2 | Corn-senesced-green-weeds (1343) | c2 | Corn-notill (1434) |
| c3 | Lettuce-romaine-4wk (616) | c3 | Corn-min (834) |
| c4 | Lettuce-romaine-5wk (1525) | c4 | Corn (234) |
| c5 | Lettuce-romaine-6wk (674) | c5 | Grass-pasture (497) |
| c6 | Lettuce-romaine-7wk (799) | c6 | Grass-trees (747) |
| | | c7 | Grass-mowed (26) |
| | | c8 | Hay-windrowed (489) |
| | | c9 | Oats (20) |
| | | c10 | Soybean-notill (968) |
| | | c11 | Soybean-min (2468) |
| | | c12 | Soybean-heavy till (614) |
| | | c13 | Wheat (212) |
| | | c14 | Woods (1294) |
| | | c15 | Bldg-grass-tree-dr (380) |
| | | c16 | Stone-steel towers (95) |

spatial resolution (3.7-meter pixels). The area covered comprises 512 lines by 217 samples. The 20 water absorption bands were discarded, *i.e.* bands: [108-112], [154-167], 224. It includes vegetables, bare soils, and vineyard fields. Salinas ground truth contains 16 classes.

## VIII. EXPERIMENTAL RESULTS

MAFE evaluation on lower dimensional representations consists of various comparisons with popular techniques on all four image scenes. The comparison includes experimental results obtained with SNE, tSNE, Isomap and LE, taking as input the Gaussian kernel neighborhood graphs obtained from a principal component analysis step that reduces the original dimension to $40$. In constructing the high dimensional neighborhood graph, we vary the number of neighbors from $k = 1$ to $k = 50$ and pick an optimal value of $k = 15$. The optimal value for $k$ ensures that the embeddings are neither too noisy and unstable nor does the geometry of

TABLE III

GROUND TRUTH CLASSES FOR IMAGE SCENES WITH THEIR RESPECTIVE SAMPLES NUMBERS

| | Botswana Hyperion | | Kennedy Space Center |
|---|---|---|---|
| c1 | Water (158) | c1 | Scrub (761) |
| c2 | Floodplain (228) | c2 | Willow swamp (243) |
| c3 | Riparian (237) | c3 | Cabbage hamm (256) |
| c4 | Firescar (178) | c4 | Cabbage palm(252) |
| c5 | Island interior (183) | c5 | Slash pine (161) |
| c6 | Woodlands (199) | c6 | Oak (229) |
| c7 | Savanna (162) | c7 | Hardwood swamp (105) |
| c8 | Short mopane (124) | c8 | Graminoid marsh (431) |
| c9 | Exposed soils (111) | c9 | Spartina marsh (520) |
| | | c10 | Cattail marsh (404) |
| | | c11 | Salt marsh (419) |
| | | c12 | Mud flats (503) |
| | | c13 | Water (927) |

the observation collapse the coordinates of dissimilar observations. All existing models are implemented as in [16], with the perplexity of the conditional distribution induced by the Gaussian kernel determined as $2^H$, where $H$ is the entropy. The parameterized bilateral kernel function presented in Section V-A, is used to generate the input graphs for MAFE with no principal component analysis step. MAFE-BR results are generated with a setting $p = q = 2$ for (31), establishing a quadratic attraction potential and a spherical Gaussian repulsion potential. MAFE-BR's magnitude parameters are set as $\xi_a = 0.4$ and $\xi_r = 10^{-4}$. MAFE-UR results are generated with a setting $p = 2, q = 1$ and parameter values $\xi_a = 0.03$ and $\xi_r = 10^{-5}$. The proposed MAFE optimization algorithm uses the norm of the gradient as the termination condition, *i.e.* $\|\nabla U(\mathcal{Z}^{(t)})\| < \epsilon$, with $\epsilon = 10^{-5}$. SNE, tSNE and sSNE obtain solutions based on standard gradient descent algorithms whose terminations are defined by the number of iterations $T$, *i.e.* terminate when $T = 1000$ iterations. Embedding maps obtained by Isomap are based on the classical formulation that admits the closed-form solution of an eigenvector structured problem, namely picking the leading components of variation, while LE is based on picking the

trailing eigenvectors.

## A. Trajectories of Gradient Vector Field

The dynamic equation in (1) serves as an approximation to the true model that describes the central path in search for optimal data manifolds. Additional insights on the uncovering of the underlying manifolds can be obtained by observing the gradient field trajectories as each map traverses towards the minimum energy configuration state of the embedding graph. The trajectory traces in 2D space provide a visual assessment of how the formation of clusters is affected by different cost functions and their corresponding optimization schemes. The gradient vector fields to be presented were obtained from mapping fifteen hyperspectral pixels from three classes of both the Botswana and the Kennedy Space Center images. Three classes are considered for each image, *i.e.Woodlands, Firescar ,* Island Interior for Botswana and *Water, Graminoid Marsh, Cabbage Palm* for Kennedy Space Center. MAFE based trajectories are compared to gradient fields obtained with tSNE and SNE. Figure 5 and Figure 6 illustrate the formation of clusters under these iterative optimized embedding algorithms. tSNE and SNE results are shown in Figures 6(d) and 6(e) from which a high degree of oscillation and random change in gradient vector directions can be observed. These instabilities causes inefficiencies on the establishment of the equilibrium distances $\epsilon_{ij}$ between pairs of maps. In contrast, starting from the same initial maps, Figures 6(b) and 6(c) demonstrates MAFE-BR and MAFE-UR techniques with smooth trajectories and no instabilities, i.e. no random change of gradient vector directions. Such smooth gradient fields are due to two important features of the proposed MAFE framework. First, the local stochastic adaptive optimization scheme used in MAFE models retains the local gradient information from its last two sequence computations thereby establishing smooth and stable iterates, it does not depend on random jitter parameters in its update rule. Second, the careful design of pair-dependent attraction and repulsion functions, as well as the sparsity of neighborhood graphs allow for a smooth interaction between long range and short range forces on pairs of maps. The smooth interaction generates a robust motion of clusters that lead to the establishment of the minimum energy configuration state much quicker. As such, the optimization scheme used in MAFE models achieves several magnitudes of convergence speed in contrast to other algorithms as shown in Figure 16.
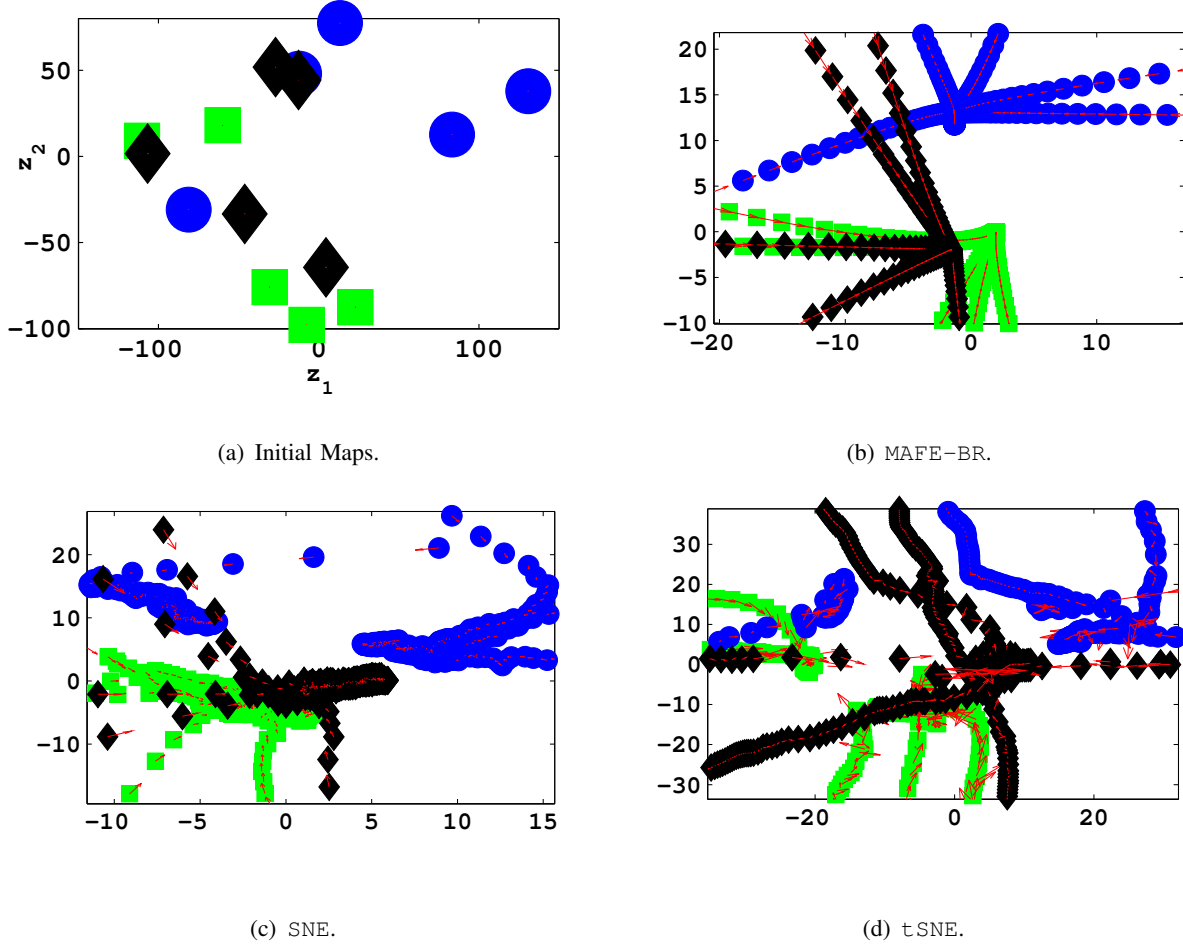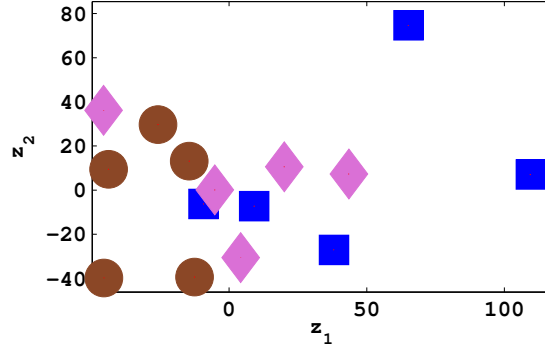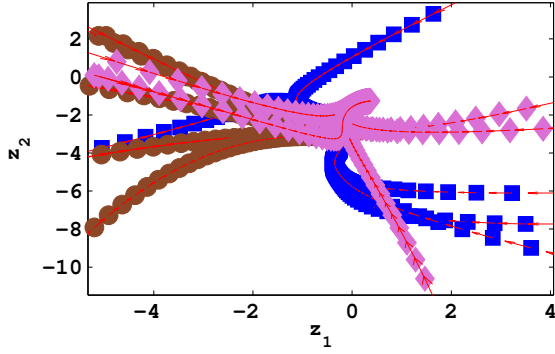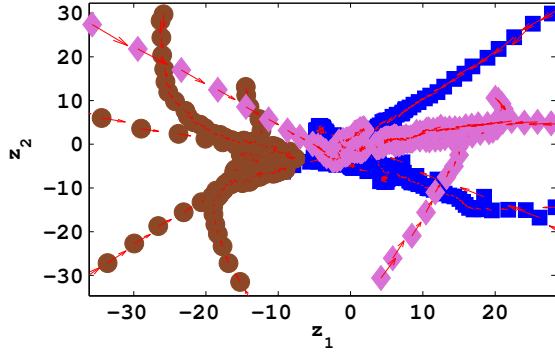
(a) Initial Maps.

(b) MAFE-BR.

(c) SNE.

(d) tSNE.

Fig. 5. Gradient based optimization illustrations over 100 iterations. MAFE-BR(b), SNE(c) , tSNE(d), are all initialized from the same seed of 15 points for Botswana maps of 3 different classes (*Woodlands, Firescar* and *Island Interior*). MAFE-BR displays smooth gradient trajectories. Similar points cluster and traverse in the direction of the gradient field as indicated by the arrows. SNE and tSNE trajectories are subject to oscillations, including collisions of maps leading to instabilities as well as the severe local maxima traps that slow down the optimization algorithm. Arrows are shown pointing in the negative direction of the gradient.

## B. Visualization of Embeddings

Nonlinear dimensionality reduction methods do tend to demonstrate better performance when embedding data sets with smaller number of classes than for problems with many classes [19]. This is caused by the complexity of data manifolds represented in problems with many disparate classes. For example, in hyperspectral imagery; similar classes of data may result in multiple manifolds due to their spatial locations, as such that presents a challenge to many existing dimensionality reduction techniques with a tendency of mapping all similar or related data onto

(a) Initial Maps.



(b) MAFE-BR.
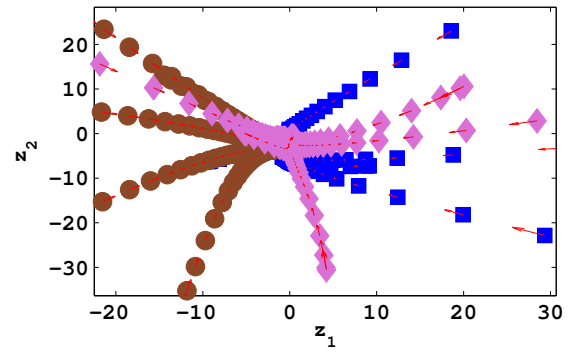


(c) MAFE-UR.



(d) SNE.
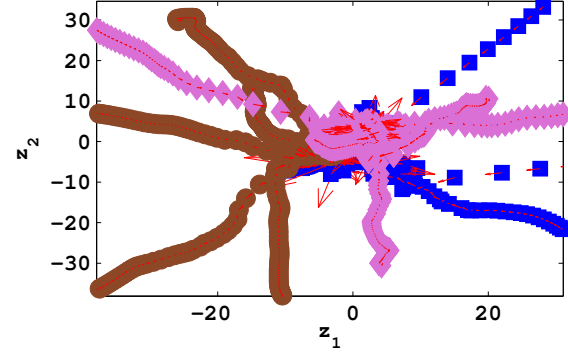


(e) tSNE.

Fig. 6.   Gradient based optimization illustrations over 100 iterations. All techniques were initialized from the same points to embed fifteen KSC pixel maps consisting of three different classes *i.e. Water, Graminoid Marsh* and *Cabbage Palm*. Both MAFE-BR and MAFE-UR demonstrate smooth gradient trajectories with points belonging to the same class clustering together and traversing in the negative direction of the gradient field ( indicated by the arrows). SNE and tSNE trajectories are subject to oscillations and instabilities leading to poor convergence of the algorithms.

a single cluster. Such characteristics increases the chances of collapsing very different classes to the same embedding representation. Using MAFE-BR and MAFE-UR models, this study makes an effort to address this challenge and provides high quality lower dimensional visualization results for complex data sets.

Figure 7 shows the Botswana RGB image and its ground references data. The embedding results obtained for MAFE-BR, SNE, tSNE, MDS, Isomap, and LE are shown in Figure 8. MAFE-BR results demonstrate different and superior embedding map compared to other methods. For example, SNE computes coordinates that seem to separate different classes well however there is a significant overlap on the *Riparian* and *Woodlands* classes. The clusters seem to be more spread implying large variances in the embedding space. MAFE-BR combined with the bilateral kernel function achieves tight spatial disjoint clusters, demonstrating no overlaps on the *Riparian* and *Woodlands* classes, the most known to be difficult classes to separate for this data set [19]. tSNE although with capability to mitigate overcrowding of points and good clustering effect, leads to significant overlap between *Riparian* and *Woodlands* classes. MDS and Isomap embeddings are similar due to the dependence on the classical MDS solution, both results show a significant overlapping of coordinates. LE produces a solution with very limited interpretation on the separation of classes.

Figure 9 shows both the RGB image and ground references of the Kennedy Space Center (KSC) data. Figure 10 shows the embedding results obtained with MAFE-BR, MAFE-UR, SNE, tSNE, Isomap, and LE. The 2-dimensional embedding results demonstrate that MAFE-BR and MAFE-UR construct very tight clusters. In addition, due to the spatially-sensitive kernel the embedding result maintains the disjoint nature observed from the land cover categories of the ground truth. Furthermore, MAFE models introduce the capability to *tile* different land cover classes when their boundaries appear to be overlapping. The models achieve better neighborhood properties as illustrated in Figure 15. SNE and tSNE results demonstrate the existence of clusters for *Water, Salt Marsh,* and *Spartina Marsh* classes. However, there is very little insight on the separability of the other ten classes. Isomap's solution demonstrates the presence of overcrowding and class overlap. LE shades very little meaningful interpretation on the structure of all classes.

Figure 11 and Figure 13 shows ground truth image data for both the Indian Pine and the Saline-A scenes. Their corresponding 2-dimensional MAFE based embeddings are shown in Figure 12 and Figure 14. The results further demonstrate quality visualizations that are generated due to
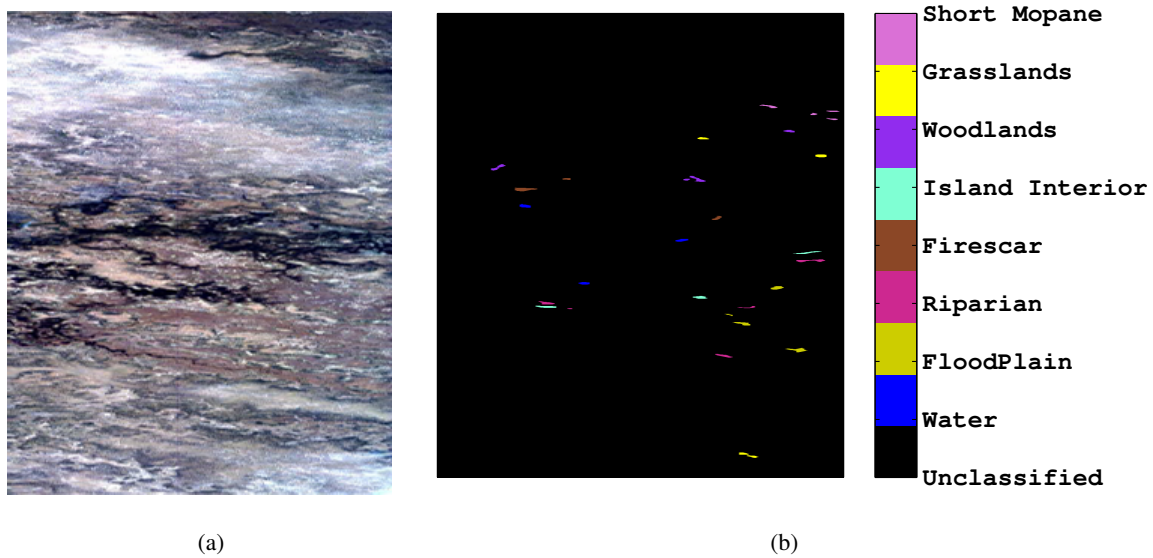
Fig. 7.    (a) RGB image of Botswana data. (b) Botswana labeled data.

use of a bilateral kernel function in combination with a MAFE based embedding framework. Use of spatial information allows for the disjoint nature of classes to be encoded in the neighborhood graph, *i.e.* it introduces the needed level of sparsity for establishing disjoint neighborhoods. The interactive fields that generates the attraction and repulsion forces in MAFE techniques enables preserving such topological relations.
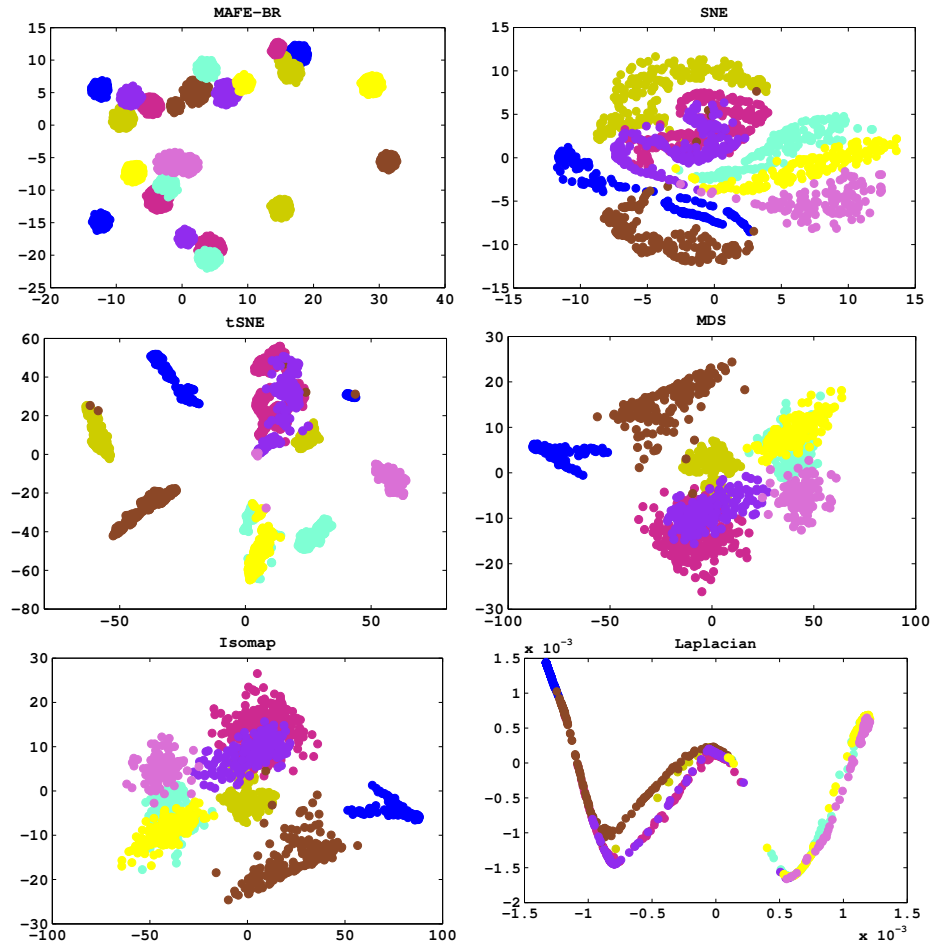
Fig. 8.    Embedding of all labeled samples of the Botswana data.

(a)                                                              (b)
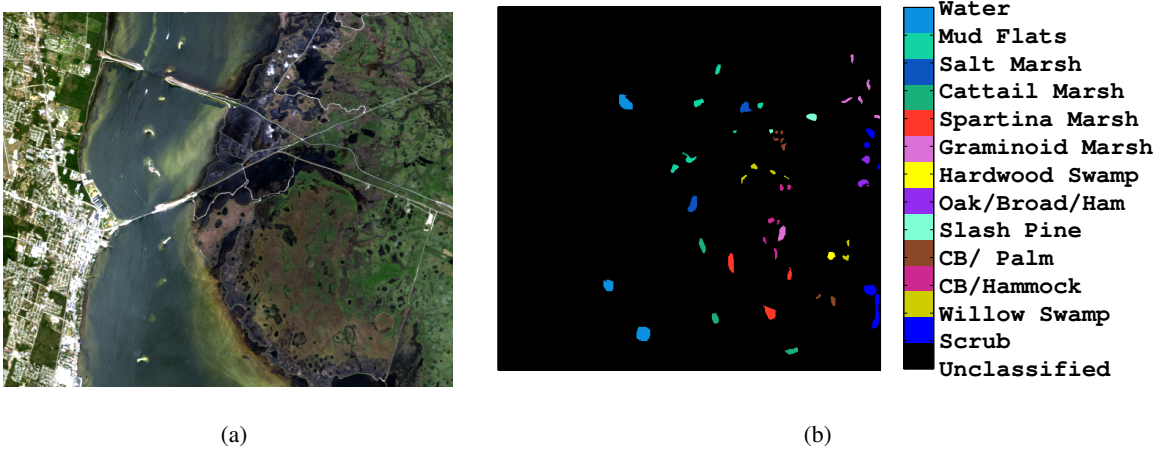
Fig. 9.    (a) RGB image of Kennedy Space Center data. (b) Kennedy Space Center labeled data.

## C. Frobenius Distance

As in many dimensionality reduction methods, the goal is to maintain the local structure of the data. Here, the study seeks to examine the local topology of manifolds obtained by MAFE models in comparison to other iterative optimized embedding algorithms *i.e.* SNE, and tSNE. The most straightforward approach to assess this is by the norm of the residual distance matrix:

$$\|\boldsymbol{D} - \hat{\boldsymbol{D}}\|_{\mathbb{F}} = \sqrt{\sum_{ij} (\hat{D}_{ij} - D_{ij})^2} \tag{35}$$

where $\|\cdot\|_{\mathbb{F}}$ denotes the Frobenius norm. It is the square root sum of the squares of elements of $\boldsymbol{D}$ and $\hat{\boldsymbol{D}}$, the high and low dimensional Euclidean distance matrices, respectively. The results shown in Figure 15 indicate that MAFE models achieve better local distance preserving representations in the sense of (35). The results also highlight the efficiency with which the optimization proposed for MAFE based models enables the discovery of neighboring related coordinates. As the number of iterations increase, SNE continues to show reduced Frobenius error due to its small magnitude repulsion which in turn has a negative effect as it leads to overcrowding of maps.

In the case of tSNE, its clear that it does not preserve local distances due to its strong repulsion forces. tSNE has a tendency of separating maps of similar instances into multiple small separated clusters, a property that might be useful for visualization and not so suitable for tasks that require keeping similar maps in close proximity, *e.g.* COIL20 data visualization and hyperspectral data
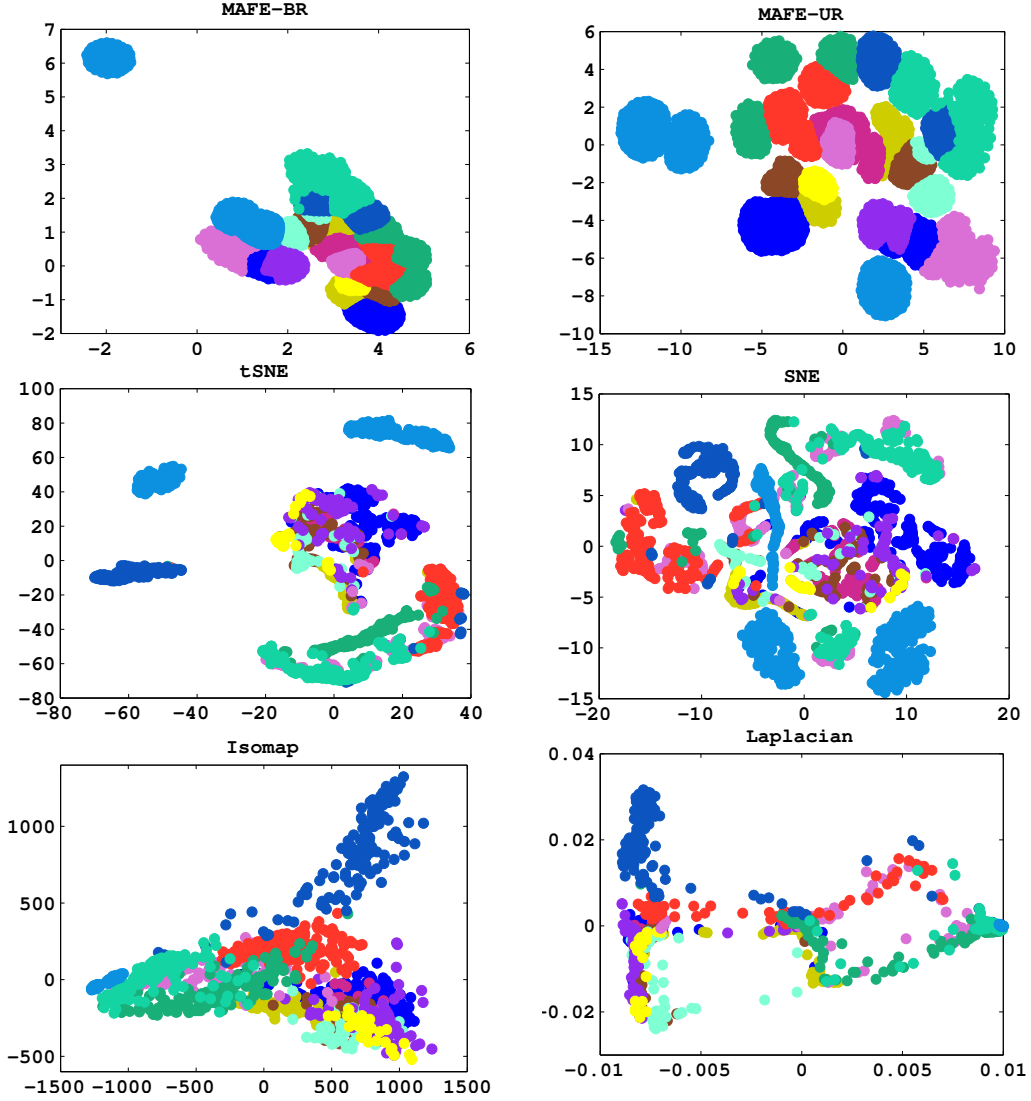
Fig. 10. Embedding of Kennedy Space Center data (showing 2600 pixels coordinates).

embedding.

## D. Semisupervised Classification

In various image analysis applications, including remote sensing, the main task entails establishing an automated image understanding process through object classification or land cover classification. Class label information is used to determine the designation of test or query samples in the lower dimensional space. Our experiments employ a one nearest neighbor(1NN) classification performance accuracy, based on the spectral angle mapper [40], [41], to determine
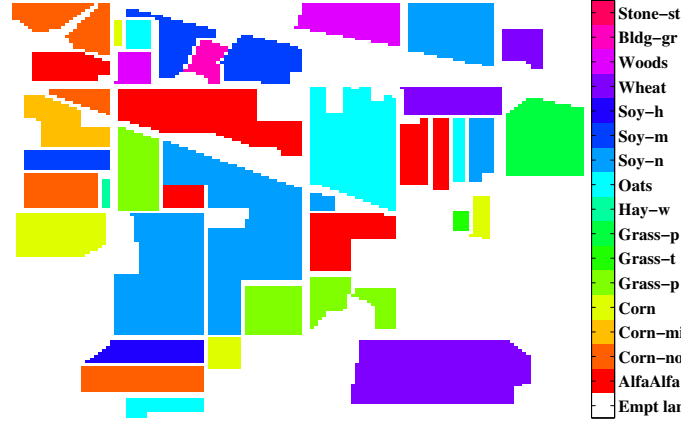
Fig. 11.   Indian Pines ground truth image.

whether the defined object classes or earth land covers, do occupy separable volumes in the lower dimensional space and if they can be discriminated successfully. The experiments are carried out in various embedding spaces. Embedded pixel maps were randomly sampled to generate $70\%$ training samples and $30\%$ testing samples, with results averaged over $10$ runs. All approaches were compared on the same data samples to maintain consistency in error comparison. The results provide several interesting observations that are consistent with the visualization maps from Chapter VIII-B.

Tables IV and V illustrates the 1NN classification performance accuracy per class. The trends observed on MAFE-BR and MAFE-UR embedding spaces indicate coordinate representations that outperform other spaces by significant margins in enabling high classification accuracy. Performance accuracies are reported by observing the mean values obtained from the Kappa statistic (KS) [42] that is computed as,

$$KS = \frac{N \sum_{i=1}^{|C|} t_{cc} - \sum_{c=1}^{|C|} t_{c+} t_{+c}}{N^2 - \sum_{c=1}^{|C|} t_{c+} t_{+c}}$$

where $N$ is the number of testing samples, $t_{cc}$ indicates the number of samples correctly classified in class $c$, $t_{c+}$ denotes the number of testing samples labeled as class $c$, and $t_{+c}$ denotes the number of samples predicted as belonging to class $c$. $|C|$ denotes the total number of classes in the data. The percentage of correctly predicted samples of the total testing samples provides the overall accuracy (OA) which is also reported in the results.
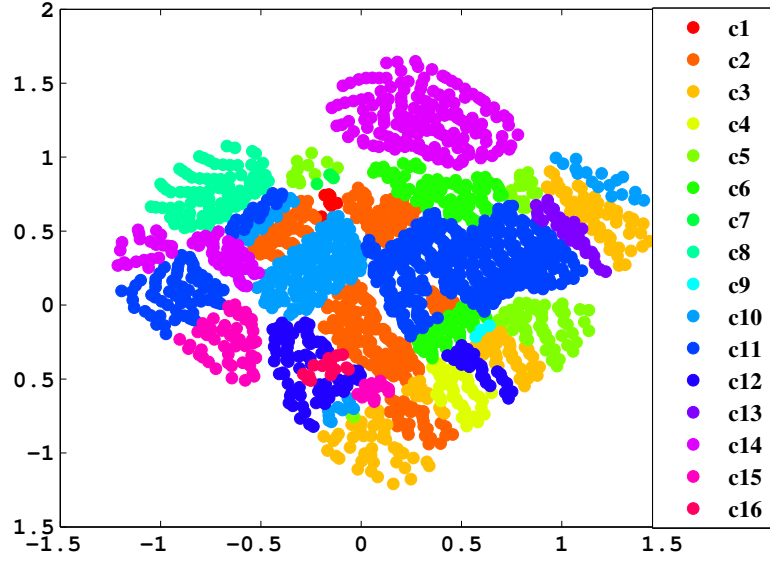
(a) MAFE-BR
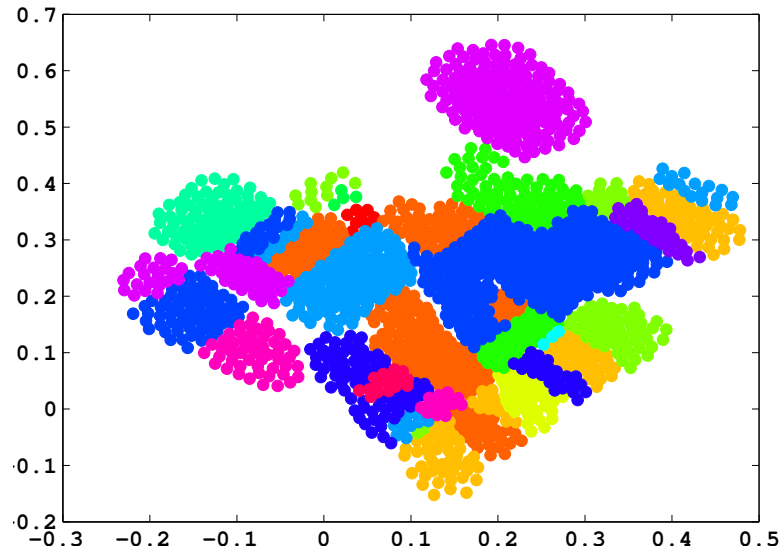


(b) MAFE-UR

Fig. 12.  Embedding of the Indian Pine Scene(showing 2500 samples to avoid clutter).

The classification results obtained on the Botswana data leads to interpretations that are consistent with the visualization analysis of Figure 8. In general, all methods seem to provide embedding coordinates that leads to reasonable classification performance accuracy. However, the lowest accuracy results are achieved with the LE embedding representations. Furthermore, the lowest accuracy per class is observed between class 3 (c3) and class 6 (c6) corresponding
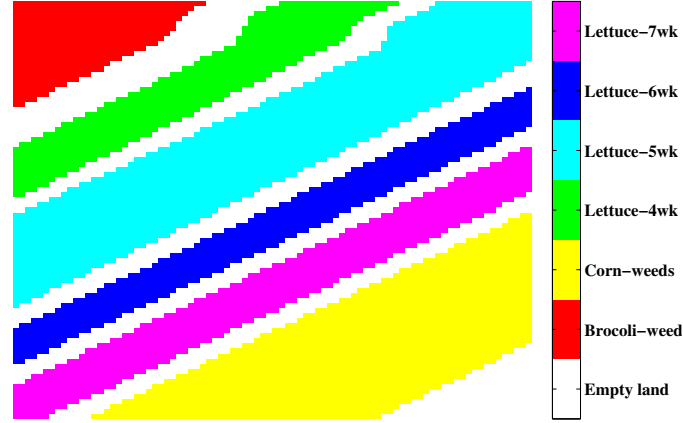
Fig. 13.   Salinas-A ground truth image.

to the *Riparian* and *Woodlands* classes, respectively. Lower classification results on c3 and c6 are expected because these two classes are the most difficult to separate, in consistency with the visual results of Figure 8 - similarly as demonstrated in [19], [43], [44]. LMNN using 1NN achieves the second best result on both data set perhaps owing to its ability to make use of class label information in learning the Mahalanobis distance metric. The objective function for LMNN does have a force field structure in which class label information is used to compute the optimal metric with a goal that k-nearest neighbors always belong to the same class (*i.e.* pulled closer by an attraction term), while example samples from other classes are separated by a large margin (*i.e.* pushed far by a repulsion term). In contrast, both MAFE-BR and MAFE-UR have objective functions that are formulated as dependent on the distance between pairs of points, and no class label information is used during computation of the maps. When classifying samples from the KSC data, a similar trend is observed with both MAFE based techniques providing a coordinate representation from which a higher 1NN classification performance is achieved. The classification results achieved for class 3(c3), class 4 (c4), class 5(c5), and class 6(c6) indicate the lowest performance in all embedding spaces except for the solution achieved by the MAFE based techniques. Based on the visualization result of Figure 10, these classes correspond to the *Cabbage Palm Hammock, Cabbage Palm/Oak Hammock, Slash Pine,* and *Oak/Broadleaf Hammock,* respectively. These are all categories of very similar upland trees. Their spectral signatures are mixed and often exhibit only subtle differences. However, the neighborhood graph
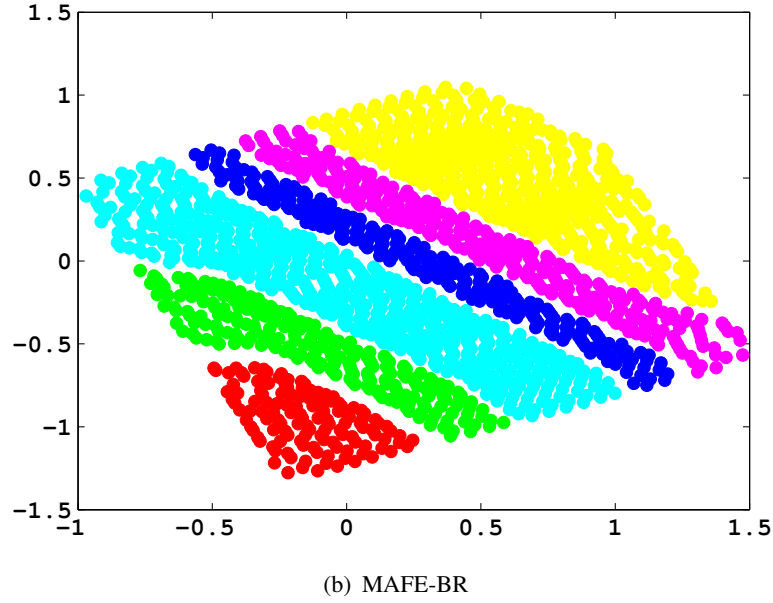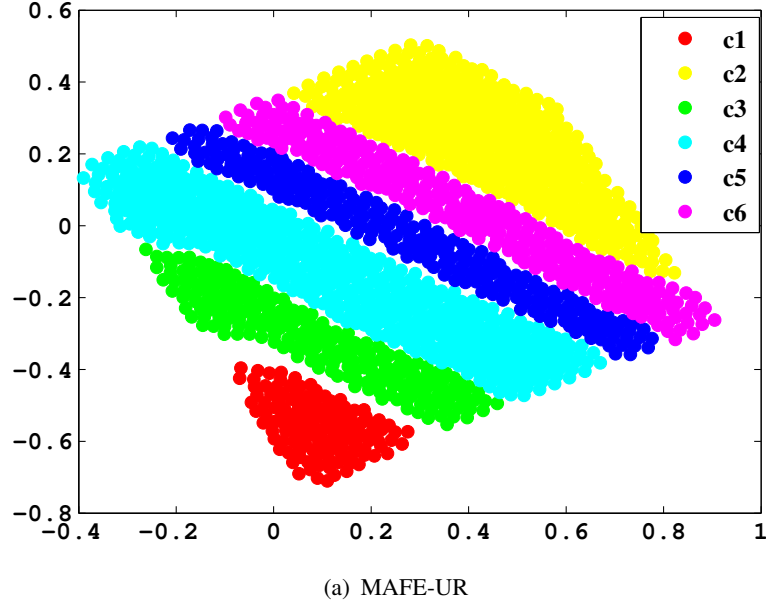
(a) MAFE-UR



(b) MAFE-BR

Fig. 14. Embedding of the Saline-A Image Scene (showing 2000 samples to avoid clutter).

structure computed from the bilateral local kernel function which incorporates spatial details plays a significant boost to the separation of different categories of objects. As such, MAFE based methods take advantage of this in addition to the smooth attraction/repulsion interaction between pairwise maps and yields embeddings that are clearly separable. On the other hand, sSNE achieves very similar performance results with its input graph structure defined to take
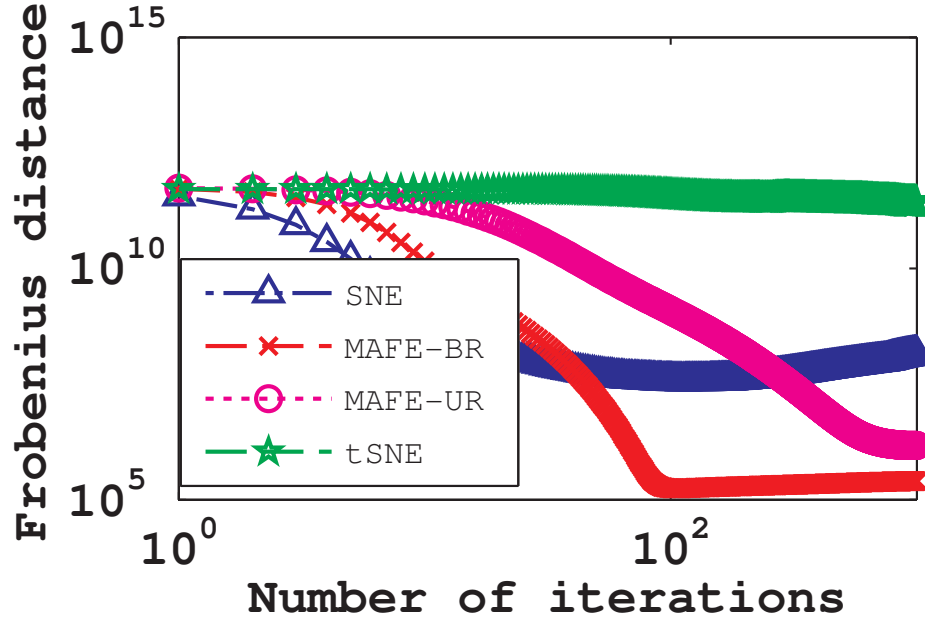
Fig. 15.   Illustration shows how the quality of local distance preservation varies with the number of iterations. The quality is defined by the Frobenius norm of the difference distance matrix , for Kennedy Space Center. MAFE related models and the SNE model show generally the same performance at preserving the local distances. The Frobenius norm results obtained from tSNE shows poor quality indicating that the method does not preserve local distance for all observed. This is due to its nature of splitting clusters of data that belong to the same class.

advantage of the spatial information in images even though it suffers from algorithmic instabilities and inefficiencies associated with the optimization of its highly nonlinear objective function.

Figure 17 shows the 1NN misclassification error plots as a function of the embedding dimension, *i.e.* $m = 1 \sim 20$. Such error plots are commonly used in dimensionality reduction algorithms that rely on the so called *manifold projection approach* for estimating the dimension of the embedding space. The manifold projection approach is used to estimate the intrinsic dimension of the embedding space by carefully predefining a higher dimensional space neighborhood graph that achieves a lower dimensional representation with better neighborhood preserving topology. Classification error plots provides one such criterion by which the intrinsic dimension of the data is chosen as the lowest dimension that allows capturing most of the variance (*i.e.* regular information) in the data with higher dimensions only adding to the redundancies. The bilateral kernel function proposed for constructing the higher dimensional neighborhood graph allows
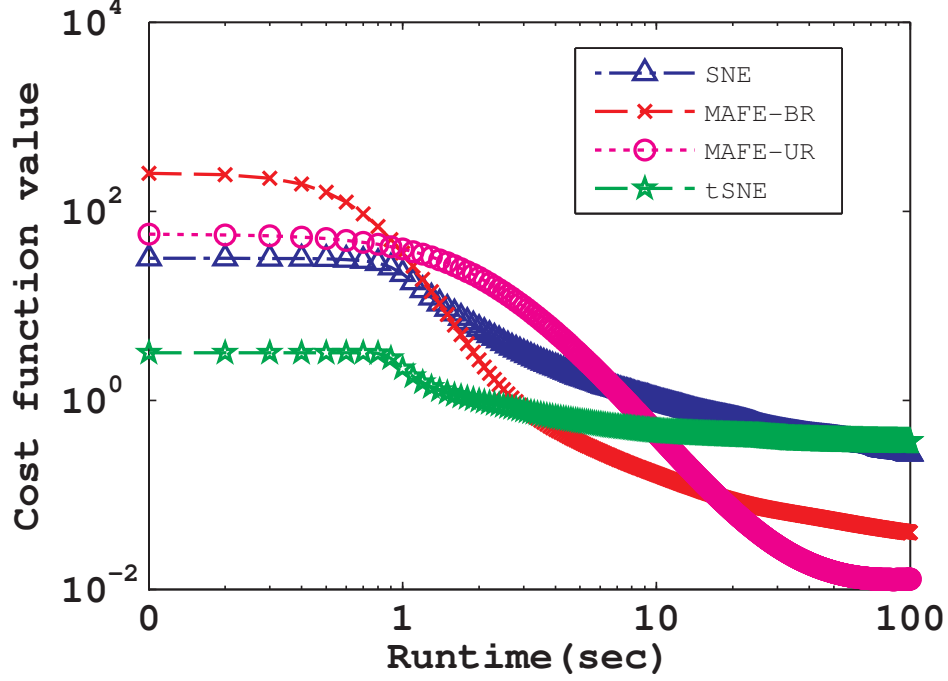
Fig. 16. MAFE, SNE and tSNE objective function optimization run times in seconds for KSC data set.

a significant automated sparsity property to be induced by estimating the covariance structure of the data. However, the study does acknowledge that choosing the intrinsic dimension is an open difficult research question that warranties a deep separate study, here we simply include an experimental approach necessary for evaluating the proposed models. From Figures 17(b) and 17(a), the optimal dimension for mapping both the $145$ dimension Botswana spectral channels and the $176$ Kennedy Space Center spectral channels is $m = 6$. The approximate optimal embedding dimension is chosen based on the smallest lowest *elbow drop* point displayed on the 1NN classification error plots in the MAFE embedding spaces.

## IX. DISCUSSION AND FUTURE WORK

MAFE nonlinear embedding techniques coupled with a bilateral kernel function demonstrated a better approach to account for nonlinear mixtures of similar categories that are captured by high resolution sensors as compared to other techniques. The MAFE general framework opens up valuable avenues coupled with a platform to derive further theoretical insights and development of new nonlinear dimensionality reduction algorithms.

TABLE IV

BOTSWANA DATA CLASSIFICATION RESULTS IN DIFFERENT EMBEDDING SPACES.

| Class | Accuracy on single class (%) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | MDS | SNE | tSNE | MAFE-BR** | MAFE-UR** | Isomap | LE | LMNN** | LFDA |
| c1 | 100 | 100 | 100 | **100** | **100** | 100 | 100 | **100** | 100 |
| c2 | 98.74 | 98.74 | 96.86 | **100** | **100** | 100 | 97.06 | **100** | 100 |
| c3* | **89.7** | **89.09** | **88.48** | 96 | **99.8** | **91.55** | **80.28** | **94.37** | **91.55** |
| c4 | 97.6 | 97.6 | 97.6 | **100** | **100** | 100 | 96.30 | **100** | 98.11 |
| c5 | 96.09 | 94.53 | 95.31 | **100** | **100** | 96.36 | 80 | **100** | 94.55 |
| c6* | **92.86** | **92.14** | **90** | **100** | **100** | **89.83** | **69.49** | **98.31** | **94.92** |
| c7 | 96.49 | 97.37 | 96.49 | **100** | **100** | 97.92 | 77.08 | **97.92** | 97.92 |
| c8 | 96.55 | 96.55 | 96.55 | **100** | **100** | 100 | 94.59 | **100** | 97.37 |
| c9 | 96.6 | 95 | 95 | **100** | **100** | 100 | 97 | **100** | 97.37 |
| KS | 95.08 | 94.75 | 93.97 | **99.33** | **100** | 95.89 | 88.45 | **98.82** | 94.55 |

* Most difficult classes to separate.

** High 1NN classification accuracy.

TABLE V
KENNEDY SPACE CENTER DATA CLASSIFICATION RESULTS IN DIFFERENT EMBEDDING SPACES.

| Class | Accuracy on single class (%) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | MDS | tSNE | MAFE-BR** | MAFE-UR** | SNE | Isomap | LE | LMNN | LFDA |
| c1 | 90.91 | 90.34 | **97.16** | **98.3** | 91.2 | 90.33 | 72 | 93.33 | 92 |
| c2 | 90.74 | 87.04 | **100** | **100** | 85.34 | 62.22 | 43.48 | 86.96 | 82.61 |
| c3* | **80** | **78.33** | **100** | **100** | 80.4 | **80.77** | **38.48** | **76.92** | **88.46** |
| c4* | **57.89** | **45.61** | **84.21** | **91.21** | 51.9 | **56** | **28** | **68** | **68** |
| c5* | **51.28** | **41.03** | **100** | **100** | 41.20 | **47.06** | **29.41** | **41.18** | **52.94** |
| c6* | **39.29** | **35.71** | **82.14** | **86.2** | **39.30** | **40** | **28** | **84** | **72** |
| c7 | 80.77 | 80.77 | **100** | **100** | 82.77 | 90.91 | 36.36 | 90.91 | 100 |
| c8 | 74.22 | 60 | **100** | **100** | 63.62 | 69.05 | 35.71 | 90.48 | 88.10 |
| c9 | 94.21 | 91.74 | **100** | **100** | 93.56 | 96.15 | 75 | 90.38 | 96.15 |
| c10 | 89.47 | 93.68 | **100** | **100** | 93.68 | 92.68 | 80.49 | 100 | 100 |
| c11 | 93.81 | 93.81 | **98.97** | **98.97** | 93.81 | 97.56 | 85.37 | 95.12 | 97.56 |
| c12 | 82.20 | 80.51 | **100** | **100** | 81.91 | 92.16 | 80.39 | 98.04 | 98.04 |
| c13 | 100 | 100 | **100** | **100** | 100 | 100 | 98.39 | 100 | 100 |
| KS | 83.04 | 80.10 | **97.86** | **99.72** | 81.90 | 86.40 | 76.71 | 92.1 | 89.2 |

* Most difficult classes to separate.

** High 1NN classification accuracy.
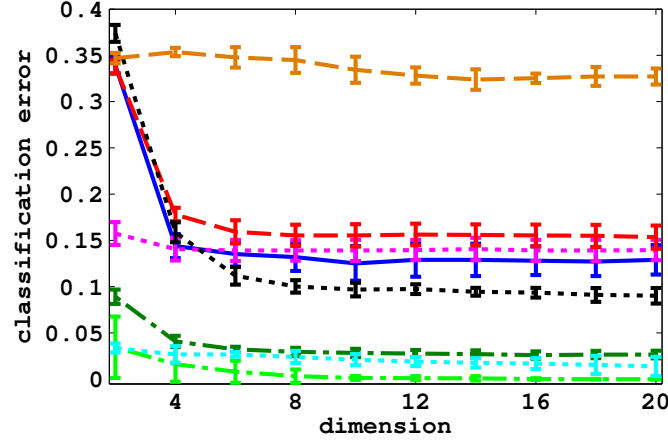
(a)
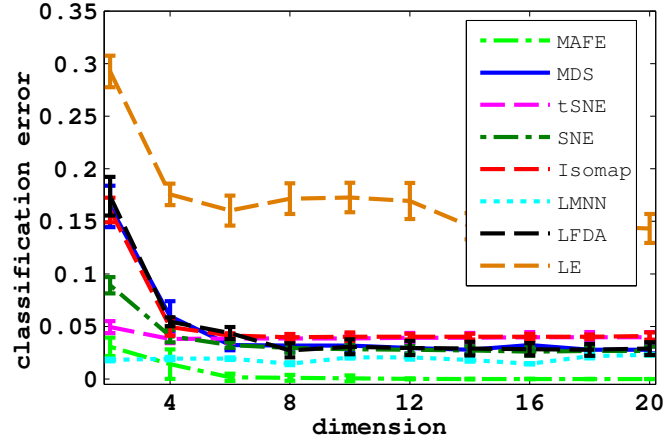


(b)

Fig. 17.   Mean $\pm$ one standard error misclassification error comparison for 1-nearest neighbor classifier based on various embedding spaces while varying the dimension. (a) Kennedy Space Center data and (b) Botswana data.

The nonlinear dimensionality reduction algorithms discussed in this study have a kernel dimension that is the square of the number of vectors in the sample space. As such the nature of their objective functions can be summarized as follows.

- For MAFE based techniques, the objective function $U : \mathbb{R}^{N \times m} \to \mathbb{R}$ is defined over the space of embedding matrices $\boldsymbol{Z} \in \mathbb{R}^{N \times m}$. The neighborhood graph's pairwise kernel similarities are represented as a $N \times N$ matrix $\boldsymbol{W}$. The complexity scales with the square of the number of observed samples, *i.e.* $O(N^2)$, both in computation and memory usage.

- For sSNE, SNE, and tSNE, the objective function $KL : \mathbb{R}^{N \times m} \to \mathbb{R}$ is defined over

the space of embedding matrices $\mathbb{R}^{N \times m}$. The probable neighbors matrix is denoted by an $N \times N$ pairwise kernel matrix $\boldsymbol{W}$. The complexity scales with the square of the number of observations, *i.e.* $O(N^2)$, both in computation and memory usage.

- For MDS, the objective function $\sigma_2 : \mathbb{R}^{N \times m} \to \mathbb{R}$ is defined over the space of matrices $\mathbb{R}^{N \times m}$. The pixel data in the problem are the geodesic distances, represented as a $N \times N$ kernel matrix $\boldsymbol{D}(\boldsymbol{Y}) = [\mathcal{D}_{\boldsymbol{Y}}(\boldsymbol{y}_i, \boldsymbol{y}_j)]$. The complexity scales with the square of the number of samples, *i.e.* $O(N^2)$, both in computation and memory usage.

## A. Algorithmic Complexity Challenges

For non-iterative methods the spectral decomposition of a large dimensional kernel encounters difficulties in at least three aspects: large memory usage, high computational complexity, and computational instability. Methods that are capable of exploiting the sparsity structure in the neighborhood graph may be useful to overcome the difficulties in memory usage and computational complexity. Approaches that incorporate the concepts of rank revealing, randomized low rank approximation algorithms, and greedy rank-revealing algorithms and randomized anisotropic transformation algorithms, which approximate leading eigenvalues and eigenvectors of dimensionality reduction kernels may lead to faster algorithms. For iterative gradient based methods, efficient second order or Newton approximation algorithms that introduce additional local information about the curvature of the objective function can be developed to speed-up the convergence to the minimum energy configuration state. However, we caution that the computation of gradient directional vectors in each iteration of the optimization scheme do continue to introduce computational hurdles for larger data sets.

As indicated, the complexity of these algorithms scales with the square of the number of pixels i.e. $O(N^2)$, a number that is prohibitive to obtain efficient solutions on large-scale (order of $10^5$ and greater) hyperspectral image pixels. This is a challenge faced by many embedding algorithms. A possible mitigating approach would be to split the pixels into overlapping patches, run a nonlinear embedding method on each patch and use the coordinates of pixels in the overlapping patches as references to merge each pair of resulting submanifolds to obtain a common global manifold. An even harder problem not addressed in this study pertains to the issue of spectral unmixing for pixels that contain more than a single land cover category [45]. If the spatial resolution of sensors is poor, increased overlap of different spectral signatures is imminent for

different land cover objects. In such cases, knowledge of the observations should be used to derive kernel representations that can incorporate such characteristics.

### B. Objective Function Formulation Challenges

The formulation of MAFE models does encounter more challenges on the design of the total energy function. For example, the appealing force field embedding intuition is subject to numerous *dynamic local maxima behaviors* during cluster formation. This can be explained from observing that the formation of clusters creates local repulsions leading to local traps for maps that still need to move closer to their most similar neighbors (as determined by the neighborhood graph weights). A simple illustration of the dynamic local maxima behavior is shown in Figure 18. The behavior is exacerbated by a weak choice on the attraction potential energy function which may lead to less effective pulling of trapped maps to their closest similar neighbors. This shortcoming does affect the convergence of the iterative embedding algorithms. The local maxima generated traps can perhaps be better handled by piecewise attraction potential functions whose short range and long range capability vary at different distances in the hope of increasing the pulling force's magnitudes to overcome the local repulsion forces from dissimilar neighbors.
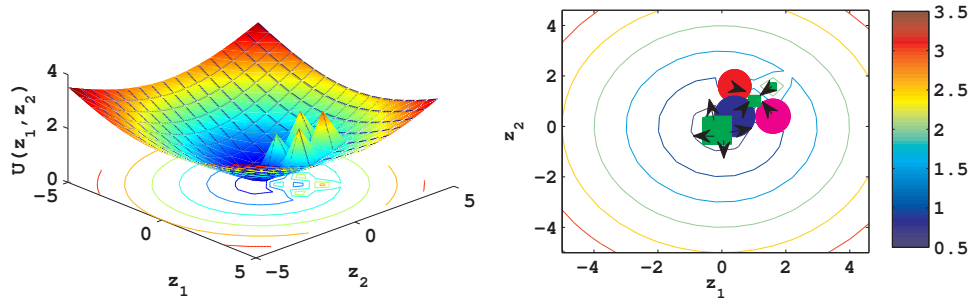


Fig. 18. Illustration of dynamic local maxima traps from a MAFE based model. Showing the 2D surface and contour plots. Arrows indicate the repulsive force fields that form during clustering. The disjoint cluster formation caused by the local repulsions on the green color class are shown on the right.(Best viewed in color.)

## X. CONCLUSIONS

This report provides design and algorithmic insights leading to new nonlinear dimensionality reduction models with strong connections to widely used methods. It introduces a novel parameterized joint spatial and photometric distance based bilateral kernel function that improves

the capability of capturing regularities within hyperspectral image pixels. Much of the disjoint relations is encoded into the neighborhood graph through spatial information and estimation of the covariance structure using the sparse matrix transform. The study then adapts a graph embedding framework, namely the multidimensional artificial field embedding, featuring the example bounded and the unbounded repulsion models, and demonstrates the general applicability of the force field intuition to problems in signal and image processing. The idea is to envision the motion of positional maps (representing graph vertices) as motivated by attraction and repulsion forces that enable a long range distance pulling of similar maps and a short range distance repulsion for all maps, respectively. The force field interpretation allows a natural way to capture this imagination and promotes the design of pair-wise interactions functions that act on pixel samples to establish the direction and magnitude of the interaction vectors. An adaptive iterative gradient based algorithm based on this notion was further implemented to yield the minimum energy configuration of the neighborhood graph. The proposed MAFE-UR and MAFE-BR models were applied to data sets acquired from remote sensing. The new algorithms proposed in the study are shown to have desirable properties that preserves the local topology of observations while inducing strong natural global structures, *e.g.* disjoint spatially motivated clusters in hyperspectral imagery. In its general form, the framework yields formulations of current popular dimensionality reduction methods with very few assumptions. Experimental work conducted on visualization, gradient field trajectories and semisupervised classification tasks demonstrates that both MAFE-UR and MAFE-BR do overcome the crowding problem and lead to better classification performance in comparison to other approaches.

## REFERENCES

[1] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2:559–572, 1901.

[2] I. Joliffe. *Principal Component Analysis*. Springer-Verlag, 1986.

[3] A. M. Martinez and A. Kak. Pca versus lda. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(2):228–223, 2001.

[4] W. Torgerson. Multidimensional Scaling I: Theory and method. *Psychometrika*, 17, 1952.

[5] M. Sugiyama. Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. *Journal of Machine Learning Research*, 8:1027–1061, 2007.

[6] M. Sugiyama, T. Idé, S. Nakajima, and J. Sese. Semi-supervised local fisher discriminant analysis for dimensionality reduction. *Machine Learning*, 78(1-2):35–61, 2010.

[7] J. A. Lee and M. Verleysen. *Nonlinear dimensionality reduction*. Springer, 1986.

[8] L. Song, A.J. Smola, K. Borgwardt, and A. Gretton. Colored maximum variance unfolding. *Adv. in Neural Information Processing Systems*, 21, 2007.

[9] S. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

[10] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comp.*, 15(6):1373–1396, 2003.

[11] V. de Silva and J. B. Tenenbaum. Global versus local methods in nonlinear dimensionality reduction. *Adv. in Neural Information Processing Sys.*, pages 721–728, 2003.

[12] M. M. Bronstein A. M. Bronstein and R. Kimmel. *Numerical Geometry of Non-Rigid Shapes*. Springer, NY, 2008.

[13] T. Kohonen and T. Honkela. *Kohonen network*. Scholarpedia, 2007.

[14] S. Yan, D. Xu, H.J. Zhang, Q. Yang, and S. Lin. Graph embedding and extensions: a general framework for dimensionality reduction. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(1):40–51, 2007.

[15] G. Hinton and S. Roweis. Stochastic neighbor embedding. *ICML*, 15:833–840, 2002.

[16] L. van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[17] M. A. Carreira-Perpinan. The elastic embedding algorithm for dimensionality reduction. *International conference on machine learning*, pages 167–174, 2010.

[18] D. Lunga and O. Ersoy. Spherical stochastic neighbor embedding of hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing*, in press, 2012.

[19] M. M. Crawford, W. Kim, and M. Li. Exploring nonlinear manifold learning for classification of hyperspectral data. *IEEE Trans. on Opt. Remote Sensing*, 3:207–234, 2011.

[20] A. E. Parr. A contribution to the theoretical analysis of the schooling behavior of fishes. *Ocassional papers of the Bingham oceanographic collection 1*, 1(1):1–32, 1927.

[21] C. M. Breder. Studies on the structure of fish schools. *Bull. Am. Mus. Nat. Hist.*, 98(1):3–27, 1951.

[22] C. Breder. Equations descriptive of the fish schools and other animal aggregations. *Ecology*, 35(3):361–370, 1954.

[23] V. Gazi and K. M. Passino. Stability analysis of social foraging swarms. *IEEE Trans. on Systems, Man, and Cybernetics*, 34(1):539–557, 2004.

[24] J.C. Latombe. *Robotic Motion Planning*. Kluwer Academic Publishers, Boston, MA, US, 1991.

[25] D. Grunbaum. Schooling as a stretegy for taxis in a noisy environment. *Animal Groups in Three Dimensions*, pages 257–281, 1997.

[26] D. GrUnbaum. Schooling as a strategy for taxis in a noisy environment. *Evolutionary Ecology*, 12:503–522, 1998.

[27] Z. Xue, Z. Liu, C. Feng, and J. Zeng. Stability analysis of exponential type stochastic swarms with time-delay. *Int. Journ. of Innovative Managemnt, Information and Production*, 2(3):1–12, 2011.

[28] S. Boyd and L. Vandenberghe. *Convex Optimization*. Camb. Uni. Press, New York, NY, USA, 2004.

[29] D. Luo, C. Ding, F. Nie, and H. Huang. Cauchy graph embedding. *In Proc. of ICML*, 2011.

[30] J. B. Tenenbaum, V. de Silva, and J.C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.

[31] R. Durrett. *Brownian Motion And Martingales in Analysis*. Wadsworth Advanced Books and Software, Belmont, California, 1951.

[32] P. Honghong and R. Raghuveer. Hyperspectral image enhancement with vector bilateral filtering. *IEEE intern. conference on Image processing*, pages 3669–3672, 2009.

[33] K. Kotwal and S. Chaudhuri. Visualization of hyperspectral images using bilateral filtering. *IEEE Transactions on Geoscience and Remote Sensing*, 48(5):2308–2316, 2010.

[34] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. *IEEE Int. Conf. on Computer Vision*, pages 839–846, 1998.

[35] S. Velasco-Forero and V. Manian. Improving hyperspectral image classification using spatial preprocessing. *IEEE TGRS*, 6(2):297–301, 2009.

[36] J. Theiler, G. Cao, L. Bachega, and C. Bouman. Sparse matrix transform for hyperspectral image processing. *Journ. of Selected Topics in Signal Proc.*, 5(3):424–437, 2011.

[37] V. P. Plagianakos, G. D. Magoulas, and M. N. Vrahatis. Ch. 2 learning rate adaptation in stochastic grdient descent. *Information Sys. Journ., Kluwer Academic Pub*, pages 15–26, 2001.

[38] J.P. LaSalle. Some extensions of Liapunov's second method. *IRE Trans. on Circuit Theory*, pages 520–527, 1960.

[39] A. L. Neuenschwander. *Remote sensing of vegetation dynamics in response to flooding and fire in the Okavango Delta-Botswana*. Ph.D. dissertation, Univ. Texas Austin, Austin, TX, 2007.

[40] A. F. H. Goetz R. H. Yuhas and J. W. Boardman. Discrimination amoung semi-arid landscape endmembers using the spectral angle members (sam) algorithm. *In Annual JPL Airborne Geoscience Workshop*, pages 147–149, 1992.

[41] K. B. Heidebrecht F. A. Kruse, A. B. Boardman and P. J. Barloon. The spectral image processing system(sips)-interactive visualization and analysis of imaging spectrometer data. *Remote Sensing Environments*, 44(2):145–163, 1993.

[42] J. Cohen. A coefficient of agreement from nominal scales. *Education Psychological Measurement*, 20:37—46, 1960.

[43] W. Di and M. M. Crawford. Active learning via multi-view and local proximity co-regularization for hyperspectral image classification. *IEEE Journal of Selected Topics in Signal. Proc.*, 5(3):618–628, 2011.

[44] L. Ma and M. M. Crawford and J. Tian. Local manifold learning-based k-nearest-neighbor for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 48(11):4099–4109, 2010.

[45] J. A. Benediktsson A. Villa, J. Chanussot and C. Jutten. Spectral unmixing for the classification of hyperspectral images at a finer spatial resolution. *IEEE Journal of Selected Topics in Signal Processing*, 5(3):521–533, 2011.